Non- and Semi-parametric Techniques for Handling Missing Data

by

Niel Hens

Promotor: Prof. dr. Marc Aerts, Copromotor: Prof. dr. Geert Molenberghs

Center for Statistics, Limburgs Universitair Centrum

Missing data arise in various settings, including surveys, clinical trials and epidemiological studies. With or without missing data, the goal of a statistical analysis is to make valid and efficient inferences about a population of interest. The issue of missing values complicates this process. Early on, modelling incomplete data relied on the use of parametric models. Recently, there is a general trend towards non- and semi-parametric approaches to relax assumptions on which parametric models typically rely. Non- and semi-parametric procedures in general will not be as efficient as model-based techniques when there is a posited model, and the model is appropriate. However, if the assumed model is not the correct one, inferences can be worse than useless, leading to misleading interpretations of the data.

In this work, a variety of non- and semi-parametric techniques are used to handle missing data problems. The material presented clearly shows the benefits of relaxing assumptions.

While starting off with a basic introduction into the field of missing data and non- and semi-parametric techniques, the successive parts of this work focus on different topics. A first part describes a kernel based imputation procedure which makes use of a non-parametric regression relationship between a partially observed response and fully observed covariate. The approach is related to the approximate Bayesian bootstrap method and can be seen as an extension of the local single imputation of Cheng (1994) to a proper local multiple imputation approach. An essential ingredient of the algorithm is the local generation of responses.

In a regression analysis, selecting an appropriate model from a candidate set of models is based on, e.g., the Akaike Information Criterion (AIC, Akaike, 1973). If however observations are incomplete, the use of complete cases can lead to wrong model choices. In a second part, two modifications of the AIC-criterion are proposed. Firstly, inverse probability weighting is used to improve upon model selection. The method is applicable to both incomplete data and design-based samples. If the weights are unknown, they are estimated using generalized additive models with penalized regression splines. Whenever only a few complete cases are available by deleting every observation with at least one missing value, weighting is not adequate anymore and imputation can provide a solution. Therefore, secondly focus is on an imputation-based AIC-criterion where imputation is non-parametric in nature by using generalized additive models with penalized regression splines. The simulations also reveal potential benefits of model selection after smoothing for fully observed regression data. The use of these AIC versions is illustrated on a case study and contrasted with tree-based methods who deal with both missing values and design.

From the existing material to deal with dropout in longitudinal studies, it is clear that a sensitivity analysis should be part of any statistical analysis. Next to providing an overview of existing sensitivity tools, the third part of the thesis describes a non-parametric sensitivity tool called 'kernel weighted influence'. It uses a 'kernel based neighbourhood' concept to explore the global and local influence towards non-random missingness for types of observations instead of observations itself in a selection model framework (Diggle and Kenward, 1994). These sensitivity tools pick up a lot of different anomalies in the data, not only deviations from the MAR-assumption. A method to oppose missing at random versus missing not at random in a selection model framework is the likelihood ratio test. The bootstrap will be used in an attempt to generate the null distribution of the likelihood ratio test statistic opposing missing not at random versus missing at random in a selection model framework.

In a last part, generalized estimating equations are used to determine the force of infection for binary clustered data. The impact of missing data on the analysis is illustrated and inverse probability weighted estimating equations are proposed. The weights are estimated non-parametrically by a generalized additive model with penalized regression splines. Several other complications in the dataset are dealt with, including the constraint for the age-specific seroprevalence to be monotone increasing. Deriving confidence intervals under these constraints is done using the bootstrap. The application of these techniques in the context of veterinary epidemiology is new and therefore considered to be a motivation for interdisciplinary collaboration between statisticians and veterinary epidemiologists working in this field.