Multilevel logistic modelling: Issues when working with large datasets

Dr. E. M. Y. Cheng*
Faculty of Medicine, University of Southampton, Southampton, United Kingdom
m.y.cheng@southampton.ac.uk

Scott Harris
Faculty of Medicine, University of Southampton, Southampton, United Kingdom
sharris@southampton.ac.uk

In public health research it is becoming increasingly common for studies to combine data that is collected at the individual level with higher-level aggregated data which could come from hospitals, specialist treatment centres or GP practices. Multilevel models are often used to analyse such datasets as they allow the hierarchical structure of the data to be taken into account. With large datasets this can result in some computational issues depending on the software package being used, the computing environment and the complexity of model being fitted. For this comparison we will focus on a two-level model, which is the most commonly seen in practice. An example dataset containing in excess of 8 million cases and a higher level term with over 32,000 levels will be used as a basis for the statistical modelling. Random subsets of varying size will be taken from this dataset and models with differing levels of complexity will be applied to them all. Variations on the number of classes in the higher-level variable will also be examined and the impacts on model fitting will be noted.

Key Words: Hierarchical models, Computer intensive methods, Software comparison.