On modelling correlated binary co-morbidities

Susana Conde Llinares[1] and Gilbert MacKenzie[1]

[1] Centre of Biostatistics, Department of Mathematics & Statistics, University of Limerick, Ireland

Abstract

A log-linear model is adopted when we analyse contingency tables that arise from p multivariate binary co-morbidities [1]. Consider the p-dimensional contingency table with exactly $n = 2^p$ cells. Let y_j be the observed count in the j^{th} cell, $j = 1, \ldots, n$; the cells are ordered lexicographically in Fortran major order. Likewise, we have the bijective mapping $j \mapsto (i_1, \ldots, i_p)$ with each $i_1, \ldots, i_p$ taking the value 0 (absent) or 1 (present) [3].
Accordingly, our model is then:

$$E(Y_j) = \mu_j = \exp(\mathrm{a}_j'\theta)$$

where Y_j is the random variable denoting the number in the j^{th} cell; a_j' is the j^{th} row of the $(n \times n)$ saturated design matrix A; and θ is the $(n \times 1)$ vector of unknown parameters measuring the influence of the constant, main effects and interactions on the response.
The development of new search algorithms in R to identify *best* fitting models efficiently is our aim; especially in high dimensional and sparse contingency tables and within the class of hierarchical log-linear models (HLLMs). We have already described the construction of a backwards elimination search algorithm (BE) which mimics existing procedures in SPSS [1]. In this paper, details of three new search algorithms, two of them completed, will be given. First, another backwards elimination search algorithm (BE2) is described. It starts with the model fitting the $m \leq p$-way interactions such that this model fitted the data and the model with all the $(m-1)$-way interactions did not; thus bounding the *best* model above and below and thereby reducing the dimension of the model search space. Second, we describe a forward selection [2] algorithm (FS) which starts with the null model and adds one effect at a time until a model that fits the data is found. These algorithms work by eliminating (BE2), or by adding (FS), one effect at a time. Then the resulting model is compared with the previous model. Ultimately, another algorithm which uses tests of partial associations (FS2) is described. Now the saturated model is always the basis in the comparisons.
The evaluation of the performance of the new algorithms by means of a comprehensive simulation study is described in detail. The comparison of the results with some standard analytical procedures will be presented. We also discuss the role of likelihood ratio tests and the analysis of residuals in model selection noting *en passant* that one model selection criterion is not necessarily a surrogate for the other. Finally, alternative model selection strategies are discussed.

References

[1] Conde Llinares S and MacKenzie G (2007) Modelling High Dimensional Sets of Binary Co-morbidities. Proceedings of the 22nd International Workshop on Statistical Modelling, pp 177-180, Barcelona, Spain

[2] Goodman, Leo A (1971) The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications. Technometrics **13**, 1:33-61

[3] O'Flaherty, M and MacKenzie, G (1982) Direct Simulation of Nested Fortran DO-LOOPS. Algorithm AS **31**, 1:71-74