Student Oral Presentation

Modelling High Dimensional Sets of Binary Co-morbidities

Susana Conde[1] and Gilbert MacKenzie[1]

[1] Centre of Biostatistics, Department of Mathematics & Statistics, The University of Limerick, Limerick, Ireland.
email: susana.conde@ul.ie & gilbert.mackenzie@ul.ie

Abstract: The construction of classical co-morbidity indices is described. When the co-morbidities are binary we advocate the use of log-linear models which better capture the dependence structure in the data. We use R to implement new search strategies which enable us to analyse, sparse, high dimensional contingency tables rapidly and hence identify the best fitting models. We apply our new algorithms to a set of real medical data.

Keywords: Co-morbidity index; binary data; hierarchical log-linear model

1 Introduction

A co-morbidity is a coexisting (or additional) medical condition co-occurring with a primary disease of interest. In phase four studies, for example, when patients are on medication, the scientific interest is often in outcome - recurrence or death. Then, the burden of co-morbidity may be an important contributory determinant of outcome - one which is often overlooked in headline reporting attributing adverse events erroneously to the original treatment.
A number of solutions have been proposed in the medical literature. For example Charlson (1987) developed a Co-morbidity Index (a CCI) based on all patients admitted to the New York Hospital-Cornell Medical Center during a 1-month period in 1984. It comprises a linear combination of the co-morbidities with (age-adjusted) weights derived from a multivariate proportional hazards model of mortality. More recently Davis (1996) working with patients on dialysis derived another score based on clinical insight into the role of co-morbidity.
The construction of such indices (or so-called *risk-scores*) by divers methods is common in the medical literature and a fundamental concern is the optimality of such techniques. Below, we criticise classical methods of CCI construction and propose alternative methods of analysing multivariate binary co-morbidities, especially when p is large.

2 Classical Indices

We define a co-morbidity index as $I = w'X$ where $w' := (w_1, w_2, \ldots, w_p)$ is the weight vector and X is the corresponding co-morbidity vector. The expected value of I is $E(I) = w'E(X)$ where, for binary co-morbidities, $E(X) = [Pr(X_1 = 1), \ldots, Pr(X_p = 1)]'$. The variance is $V(I) = w'\Sigma w$ with $\Sigma = V(X)$, where (r_{th}, s_{th}) element is $\sigma_{rs} = Pr(X_r = 1 \cap X_s = 1) - Pr(X_r = 1)Pr(X_s = 1)$. An interesting case is $w = 1$ ie, the unit vector, whence $I = w'X$ is a simple count of the co-morbidities, it being assumed clinically (and erroneously in many cases) that the risk of outcome is an increasing function of I. The assumption that $w_u > 0$, $\forall u, u = 1, \ldots, p$ can also be rather dubious in practice. A key point is, that because the variables are binary and not MV Normal, their dependence structure is not summarised appropriately in the $p \times p$ variance covariance matrix, Σ.

3 Model Formulation

Given p binary co-morbidities we consider a p-dimensional contingency table with exactly $n = 2^p$ cells. Let n_j be the observed frequency (the count) in the jth. cell, $j = 1, \ldots, n$, where the cells are ordered lexicographically in Fortran major order and 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), MacKenzie & O'Flaherty (1982). Then our basic model is the usual log-linear model for contingency tables in which:

$$E(N_j) = \mu_j = \exp(\mathrm{a_j'}\theta) \tag{1}$$

where N_j is the random variable denoting the number in the jth. cell, $\mathrm{a_j'}$ is the jth. 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. From the last equation we have:

$$\log \mu_j = \mathrm{a_j'}\theta = \alpha_0 + \alpha_{1\mathrm{i}_1} + \alpha_{2\mathrm{i}_2} + \ldots + \alpha_{\mathrm{p}\mathrm{i}_\mathrm{p}} + (\alpha_{1\mathrm{i}_1}\alpha_{2\mathrm{i}_2}) + (\alpha_{1\mathrm{i}_1}\alpha_{3\mathrm{i}_3}) + \ldots + (\alpha_{1\mathrm{i}_1}\alpha_{2\mathrm{i}_2}\alpha_{3\mathrm{i}_3}) + \ldots + (\alpha_{1\mathrm{i}_1}\alpha_{2\mathrm{i}_2}\ldots\alpha_{\mathrm{p}\mathrm{i}_\mathrm{p}})$$

For inference we use the conditional Poisson model (Birch, 1963), so that $Pr(N_j = n_j) = exp\{-\mu_j\}\mu_j^{n_j}/n_j!$, leading to:

$$\ell(\theta) \propto \sum_{j=1}^{k} [-\exp(a_j'\theta) + n_j a_j'\theta] \tag{2}$$

$$i_{r,s}(\theta) = \frac{\partial^2}{\partial\theta_r \partial\theta_s}\ell(\theta) = \sum_{j=1}^{k} a_{jr}.a_{js}\exp(a_j'\theta_j) \tag{3}$$

where $1 \leq r, s \leq k$ and $k = n$ in the saturated case. We consider the class of hierarchical log-linear models (HLLMs) as a first step (Goodman, 1971).

TABLE 1. Tests of m - way effects are zero with 15 co-morbidities. The best fitting models must contain some 3-way interaction terms.

m	df	LR	P
1	15	689703.2000	0.0000
2	105	5998.8920	0.0000
3	445	480.3214	0.1198

4 Paradigms & Problems

Our adoption of this framework is predicated on the need to address some open problems in different, but related, modelling areas. For example, much original log-linear modelling was formulated in a model development environment dating back to the 1970's where $p = 10$ was considered very large. Then, today's data-mining paradigm was not envisaged and the original ideas have become ossified in legacy code in the major software packages. Accordingly, one objective of the current research is to relax these constraints by developing a new package in R. Yet another challenge is the ability to address the analysis of sparse, high-dimensional, contingency tables which might arise, for example, in thresholded micro-array data. The ability to search within these high dimensional spaces efficiently and so identify the model best supported by the data is a key objective of this research. Such searches may be facilitated by sacrificing high order interaction terms, replacing them by random effects terms instead, thereby extending the model class from a GLM to a GLMM.

5 Results

A dataset, comprising 48,158 subjects, half of whom had Chronic Obstructive Pulmonary Disease (COPD) and an equal number who were COPD-free, was analysed. A total of $p = 15$ co-morbidities were recorded. These included the presence or absence of: Myocardial Infarction, Congestive Heart Failure, Peripheral Vascular Disease, Cerebrovascular Disease, Dementia, Rheumatologic Disease, Peptic Ulcer, Mild Liver Disease, Diabetes, Hemiplegia or Paraplegia, Lung Cancer, Other Cancers, Other Respiratory Disease, Nervous System Disorder and Psychiatric Disorder.
We implemented a backwards elimination search algorithm in R, using the Iterative Proportional Fitting algorithm (Haberman, 1972) to identify the best fitting class of models. This algorithm, which tests whether the m-way interactions are exactly zero, identified the class of HLLM models including as a maximum the 3-way interactions (Table 1). This set was also identified by another algorithm which tested whether the m-way or higher order

effects were zero. One of the models involved in the comparison in the 3rd row of Table 1 contains exactly all possible 3-way interaction terms, namely 455. The best fitting model(s) in this class have yet to be identified. In total, there are $2^{455} - 1$ possible models containing at least one 3-way interaction and no higher order terms. The set of all possible 3-way interactions may now be viewed as defining another hierarchical (sub-)class of models, which can be searched (backwards or forwards) using a variant of our existing algorithms. In this way the best-fitting model(s) can be identified rapidly and compared with the results of conventional (e.g., best-subset) search strategies. At the time of writing, we are developing our new search strategies in R and will present these and other methodological innovations in the main paper.

6 Discussion

We have outlined herein the construction of conventional co-morbidity indices and highlighted some limitations of interpretation, especially in relation to dependence structures. For binary co-morbidities we propose a log-linear modelling approach which more appropriately captures the dependence between the measured co-morbidities. The method facilitates implementation in R which is free of the many restrictions imposed by existing algorithms in mainstream software packages (eg, in SPSS p=10 maximally, or $p = 8$ when generating flat contingency tables). In the R environment we have been able to develop new search strategies which allow us to identify best-fitting models efficiently.

References

Birch, M.W. (1963). Maximum Likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B*, **25**, 220-233.

Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R. (1987). , A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronical Disease*, **40**, 5, 373-383.

Goodman, L.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.

Haberman, S.J. (1972). Algorithms AS 51: Log-linear Fit for Contingency Tables. *Applied Stats.* **21**, 2, 218-225.

Harnett, P., MacKenzie, G., Al-Tawara, Y., Davies, S., Harden, P., Russell, G. and Naish, P. (2007?). *A 5 year prospective study of factors which influence selection for and survival on dialysis.* Unpublished yet.

O'Flaherty, M. and MacKenzie, G. (1982). Direct Simulation of Nested Fortran DO-LOOPS. *Statistical Algorithms. Applied Statistics*, **31**, No. 1.