Computational Statistics and Data Analysis 56 (2012) 1115–1130

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Stratified additive Poisson models: Computational methods and applications in clinical epidemiology

Ian C. Marschner [a,b,*], Alexandra C. Gillett [a], Rachel L. O'Connell [b]

[a] Department of Statistics, Macquarie University, Sydney, Australia

[b] NHMRC Clinical Trials Centre, University of Sydney, Sydney, Australia

ARTICLE INFO

Article history:
Available online 10 August 2011

Keywords:
Additive model
Clinical trial
ECM algorithm
Poisson regression
Risk factor model

ABSTRACT

Risk factor models in clinical epidemiology are important for identifying individuals at high risk of poor health outcomes and for guiding intervention strategies to reduce risk. Such models are often based on generalised linear models (GLM) with a multiplicative risk, rate or odds assumption. However, in practice some risk factors may act additively, in which case the use of a multiplicative model will lead to spurious interactions among risk factors. Computational methodology is developed for fitting non-GLM Poisson regression models that have an additive component with multiplicative stratification. These stratified additive Poisson models, which can also be applied to binomial data, provide an additive–multiplicative framework that allows greater flexibility than multiplicative models. Non-negativity constraints and high dimensionality are dealt with using an Expectation-Conditional-Maximisation (ECM) algorithm that oscillates between the multiplicative and additive components of the model. As well as providing highly stable convergence properties in a potentially unstable setting, the method allows flexible modelling features such as unspecified isotonic regression functions. The methodology is illustrated with an analysis of heart attack mortality in a large clinical trial, and it is found that the combination of additive and multiplicative components allows a more parsimonious risk factor model by removing the need for interaction terms. R code to implement the method is provided as supplementary material.

1. Introduction

An important area of clinical epidemiology is the development of statistical models that use an individual's risk factor profile to quantify their likelihood of experiencing a particular disease event (Harrell, 2001; Steyerberg, 2009). Such models are key to identifying high risk individuals and for understanding how best to intervene to reduce risk. The occurrence of disease events is often measured using either the rate of events per unit of time, or the risk (probability) of an event within a specific time period. Thus, a natural way to develop risk factor models is to fit a Poisson or binomial generalised linear model (GLM) to data on the number of disease events observed within each combination of the available risk factors.

The most common GLMs used in risk factor modelling are based on multiplicative assumptions about the risk or rate of disease. For binomial data, logistic regression has traditionally been used and corresponds to a multiplicative assumption on the risk odds. However, since risk ratios are usually more interpretable than odds ratios, the log link binomial GLM is now often preferred to the logistic model, and corresponds to a multiplicative risk assumption (McNutt et al., 2003; Speigelman and Hertzmark, 2005; Lumley et al., 2006). For Poisson data, log linear regression is most common and corresponds to a multiplicative rate assumption.

* Correspondence to: Department of Statistics, Macquarie University, NSW 2109, Australia. Tel.: +61 2 9850 8557; fax: +61 2 9850 7669.
E-mail address: ian.marschner@mq.edu.au (I.C. Marschner).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.08.002

While multiplicative models provide the mainstay of risk factor modelling, in practice it is possible that risk factors have an additive relationship with the occurrence of disease events. In this case rate differences and risk differences are better measures of disease occurrence and identity link Poisson or binomial models are often used, sometimes with simplifying approximations (Speigelman and Hertzmark, 2005; Cheung, 2007). While such additive models are less common in practice than multiplicative models, partly due to the potential for computational instability, they have been used effectively in real applications (Grotvedt et al., 2008; Gao et al., 2009).

Since both additive and multiplicative risk factor models have relevance in practice, a natural extension of these models is an additive–multiplicative model. This allows some risk factors to contribute to the model in an additive way, while others can contribute multiplicatively. However, since additive Poisson and binomial GLMs can be numerically unstable, additive–multiplicative models tend to inherit the numerical instability of their additive component. For additive binomial models the most common way of dealing with this instability is to use the additive Poisson model as a "working model" which provides unbiased estimators whose standard errors can be estimated using the so-called robust sandwich estimator (Hardin, 2003; Speigelman and Hertzmark, 2005). If the additive Poisson model is numerically unstable then recent methodology developed by Marschner (2010) can be used for reliable fitting of the model. Here we generalise this methodology to additive Poisson models having multiplicative stratification, that is, stratified additive Poisson models.

Our computations use the Expectation-Conditional-Maximisation (ECM) algorithm (Meng and Rubin, 1993; McLachlan and Krishnan, 2008) and involve oscillating between the additive and multiplicative components of the model. This ECM algorithm is applied repeatedly to maximise the Poisson log-likelihood function within a sequence of subsets of the parameter space, after which the maximum likelihood estimate (MLE) is determined from among the subset maxima. The resulting method is a numerically stable way to fit additive–multiplicative models, in both Poisson and binomial contexts, and remains reliable in high dimensions while accommodating flexible modelling features such as unspecified isotonic regression functions. The supplementary material for this paper includes source code to implement the method in R (R Development Core Team, 2011).

2. Motivating clinical application

As motivation for additive and additive–multiplicative risk factor models, we consider data on 16,949 individuals who experienced a heart attack (acute myocardial infarction) and were enrolled in the ASSENT-2 clinical trial (ASSENT-2 Investigators, 1999). This study randomised patients to one of two interventions within 6 hours of their heart attack (the drugs tenecteplase and alteplase), and recorded a primary outcome of death within 30 days of the heart attack. While the primary purpose of the study was to undertake a randomised treatment comparison, large-scale clinical cohorts such as this provide a rich source of data for the development of risk factor models that can be used to stratify heart attack cases according to their risk of mortality. In the past, risk factor models for mortality following heart attack have invariably been developed using a multiplicative assumption for the risk or the risk odds. A recurring feature of such models is the presence of interactions between risk factors, most notably between the patient's age and the severity of heart failure (Lee et al., 1995; O'Connell and Hudson, 2009). This interaction is in the direction of diminishing age-specific risk ratio (or odds ratio) for increasing severity. Similar interactions have also been observed in other cardiovascular contexts where multiplicative models are used, as discussed by Marschner et al. (2007).

While it is plausible that there is genuine effect modification occurring in cardiovascular disease, another possible explanation is that the interactions are an artefact of the multiplicative assumption. For example, if mortality following heart attack actually depends additively on age and severity, then fitting a multiplicative model would lead to an artefactual interaction in the direction described above. This can be seen by considering the behaviour of the risk ratio when the risk is an additive function of age and severity. In that case the risk ratio for age a_1 compared to age $a_2 < a_1$, for a given severity level s, is

$$\mathrm{RR}(a_1, a_2; s) = \frac{f(a_1) + g(s)}{f(a_2) + g(s)},$$

where f and g are some non-decreasing functions. This age-specific risk ratio decreases to 1 as the severity level increases, which is consistent with the direction of the interaction observed empirically when using multiplicative models. This suggests a possible alternative explanation, that the risk is additive rather than multiplicative.

Descriptive analyses of the ASSENT-2 data seem to support this explanation. Fig. 1 plots the proportion of deaths by age and severity. Here, the severity of heart failure is measured using the standard four level Killip classification, with the third and fourth levels combined due to small numbers in the latter. The results show that while there is a strong interaction on the multiplicative (log) scale there is effectively no interaction on the additive scale. In particular, the risk difference between the oldest and youngest age groups is highly stable over the three severity groups, ranging from 14.4% to 15.7%. In contrast, there is considerable variation in the risk ratio of the oldest age group compared to the youngest, ranging from 10.4 in the lowest severity group to 1.5 in the highest severity group.

The above descriptive analyses suggest that an additive risk model might be preferable to a multiplicative risk model. However, there are other potential risk factors to consider, including the patient's baseline characteristics, the type of treatment received and the context in which the patient was treated (e.g. the hospital, country or region). Such risk factors could reasonably be expected to act multiplicatively, even in the presence of additive effects from other risk factors. This