Journal of Multivariate Analysis 111 (2012) 94–109

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data

Geert Molenberghs [a,b,*], Geert Verbeke [b,a], Samuel Iddi [b,a], Clarice G.B. Demétrio [c]

[a] Biostatistical Centre, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium

[b] Center for Statistics, Universiteit Hasselt, B-3590 Diepenbeek, Belgium

[c] ESALQ, Piracicaba, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 12 September 2011
Available online 1 June 2012

AMS subject classifications:
62F99
62P10

Keywords:
Bernoulli model
Binomial model
Beta-binomial model
Conjugacy
Logistic-normal model
Maximum likelihood
Strong conjugacy

ABSTRACT

Non-Gaussian outcomes are often modeled using members of the so-called exponential family. Notorious members are the Bernoulli model for binary data, leading to logistic regression, and the Poisson model for count data, leading to Poisson regression. Two of the main reasons for extending this family are (1) the occurrence of overdispersion, meaning that the variability in the data is not adequately described by the models, which often exhibit a prescribed mean-variance link, and (2) the accommodation of hierarchical structure in the data, stemming from clustering in the data which, in turn, may result from repeatedly measuring the outcome, for various members of the same family, etc. The first issue is dealt with through a variety of overdispersion models, such as, for example, the beta-binomial model for grouped binary data and the negative-binomial model for counts. Clustering is often accommodated through the inclusion of random subject-specific effects. Though not always, one conventionally assumes such random effects to be normally distributed. While both of these phenomena may occur simultaneously, models combining them are uncommon. This paper starts from the broad class of generalized linear models accommodating overdispersion and clustering through two separate sets of random effects. We place particular emphasis on so-called conjugate random effects at the level of the mean for the first aspect and normal random effects embedded within the linear predictor for the second aspect, even though our family is more general. The binary and binomial cases are our focus. Apart from model formulation, we present an overview of estimation methods, and then settle for maximum likelihood estimation with analytic-numerical integration. The methodology is applied to two datasets of which the outcomes are binary and binomial, respectively.

1. Introduction

Like other outcome types, binary and binomial data are often measured in a longitudinal or otherwise hierarchical context. Over the last half century, a whole collection of modeling approaches has been put forward. Many are placed within the generalized linear modeling (GLM) framework [22,16,1], a unifying framework based on the so-called exponential family distributions. That said, a key feature of the GLM framework and many of the exponential family members, the so-called *mean-variance relationship*, may be overly restrictive. This relationship indicates that the variance is a deterministic function of the mean. For example, for Bernoulli outcomes with success probability $\mu = \pi$, the variance is $v(\mu) = \pi(1-\pi)$.

[*] Corresponding author at: Center for Statistics, Universiteit Hasselt, B-3590 Diepenbeek, Belgium.
E-mail address: geert.molenberghs@uhasselt.be (G. Molenberghs).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.05.005

In contrast, for continuous, normally distributed outcomes, the mean and variance are entirely separate parameters. While i.i.d. binary data cannot contradict the mean-variance relationship, i.i.d. binomial data can. Both data types are scrutinized here.

The above explains why early work has been devoted to formulating models that explicitly allow for dispersion not following the base models. It is often referred to as overdispersion, but underdispersion can occur as well. Hinde and Demétrio [9,10] provide broad overviews of approaches for dealing with overdispersion, considering moment-based as well as full-distribution avenues. For purely binary data, hierarchies need to be present in the data to violate the mean-variance link. One such class of hierarchies is with repeated measures or longitudinal data, where an outcome on a study subject is recorded repeatedly over time. Apart from the presence of extra dispersion, hierarchies in the data imply the presence of association between measurements on the same unit as well. Thus, a flexible parametric model ought to properly model the mean function, the variance function, and the association function. While the so-called generalized linear mixed model (GLMM, [5,2,30]) has become the dominant tool for hierarchical non-Gaussian data.

Molenberghs et al. [19, henceforth MVD] and Molenberghs et al. [20, henceforth MVDV] showed that accommodating either overdispersion or hierarchically-induced association may fall short of properly modeling the data. Therefore, they proposed a so-called combined modeling framework encompassing both. MVD focused on counts, whereas MVDV laid out a general framework. They briefly exemplified it, in counts, time-to-event, and binary outcomes, but did not tackle binomial outcomes. This is the subject of the current paper, with emphasis on the subtle differences between them.

The paper is organized as follows. In Section 2, two motivating case studies are presented, one exhibiting binary outcomes, the other of a binomial type. Analysis of these is relegated to Section 6. Basic ingredients for our modeling framework, standard generalized linear models, extensions for overdispersion, with particular emphasis on the beta-binomial model, and the GLMM, are the subject of Section 3. The proposed, combined model is described and further studied in Section 4. Parameter estimation is touched upon in Section 5. A simulation study, comparing the proposed model and the GLMM, is described and results presented in Section 7.

2. Case studies

2.1. Onychomycosis

These data come from a randomized, double-blind, parallel group, multicenter study for the comparison of two oral treatments (coded as *A* and *B*) for toenail dermatophyte onychomycosis (TDO), described in full detail by De Backer et al. [4] and analyzed before, among others, in [18]. TDO affects about 2% of Western populations [25]. The anti-fungal compounds studied here need to be taken during three months until the whole nail has grown out healthily. A total of 2 × 189 patients were randomized. Subjects were followed monthly during the first quarter, during which the treatment was given, and then scored once during three more quarters. Including the baseline measurement; this amounts to a maximum of seven measurements per subject. The outcome of interest here is the severity of the infection, coded as 0 (not severe) or 1 (severe) by the treating physician. The question of interest was whether the percentage of severe infections decreased over time, and whether that evolution was different for the two treatment groups.

2.2. Iron-deficient diets in rats

These data result from an experiment where female rats were put on iron-deficient diets [27]. This dataset has been analyzed by Liang and McCullagh [15] and Moore and Tsiatis [21]. In [1], the data were used to estimate several logit models. Experimental rats were divided into 4 groups, one of which is a control group. The number of female rats per group (total number of fetuses per group) are: 31 (327) for placebo, 12 (118) for low dose, 5 (58) for medium dose, and 10 (104) for high dose. Weekly injections of iron supplement were to bring the rats' iron intake to normal levels. Rats in the placebo group were given a placebo injection, the others got three different doses of the iron supplements. Rats were made pregnant and sacrificed 3 weeks later and the total number of fetuses and the number of dead fetuses in each litter were counted. Hemoglobin levels of the mothers were also measured.

3. Building blocks

In Section 3.1, we will first describe the conventional exponential family and generalized linear modeling based on it. Section 3.2 is devoted to a brief review of models for overdispersion.

3.1. Standard generalized linear models

A random variable *Y* follows an exponential family distribution, also known as exponential dispersion model [12] if the density is of the form

$$f(y) \equiv f(y|\eta,\phi) = \exp\left\{\phi^{-1}[y\eta - \psi(\eta)] + c(y,\phi)\right\}, \tag{1}$$