Computational Statistics and Data Analysis 57 (2013) 233–245

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

A hierarchical Bayesian approach for the analysis of longitudinal count data with overdispersion: A simulation study

Mehreteab Aregay[a], Ziv Shkedy[b], Geert Molenberghs[b,a,*]

[a] I-BioStat, Katholieke Universiteit Leuven, Leuven, Belgium
[b] I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium

ARTICLE INFO

Article history:
Received 2 August 2011
Received in revised form 19 June 2012
Accepted 20 June 2012
Available online 27 June 2012

Keywords:
Deviance information criteria
Hierarchical Poisson–Normal model (HPN)
Hierarchical Poisson–Normal overdispersed model (HPNOD)
Overdispersion

ABSTRACT

In sets of count data, the sample variance is often considerably larger or smaller than the sample mean, known as a problem of over- or underdispersion. The focus is on hierarchical Bayesian modeling of such longitudinal count data. Two different models are considered. The first one assumes a Poisson distribution for the count data and includes a subject-specific intercept, which is assumed to follow a normal distribution, to account for subject heterogeneity. However, such a model does not fully address the potential problem of extra-Poisson dispersion. The second model, therefore, includes also random subject and time dependent parameters, assumed to be gamma distributed for reasons of conjugacy. To compare the performance of the two models, a simulation study is conducted in which the mean squared error, relative bias, and variance of the posterior means are compared.

1. Introduction

In medical research, data are often collected in the form of counts, e.g., corresponding to the number of times that a particular event of interest occurs. A common model for count data is the Poisson model, which is rather restrictive, given that variance and mean are equal. Often, in observed count data, the sample variance is considerably larger (smaller) than the sample mean—a phenomenon called overdispersion (underdispersion). Generically, this is referred to as extra-(Poisson)-dispersion (Iddi and Molenberghs, 2012). If not appropriately accounted for, extra-dispersion may cause serious flaws in precision estimation, and inferences based there upon (Breslow, 1990). However, such excess variation has little effect on the estimation of the regression coefficients of primary interest (Cox, 1983).

One of the approaches to this problem is to assume a specific, flexible parametric distribution for the Poisson means associated with each observed count. Margolin et al. (1981) assumed a gamma mixing distribution for the Poisson means which leads to the negative binomial model. The advantage of this parametric approach is that parameter estimates may be obtained by maximum likelihood, leading to estimates that are asymptotically normal, consistent, and efficient if the parametric assumptions are accurate (Cramér, 1946; Wald, 1949).

Under conditions discussed by Cox (1983), maximum likelihood methods maintain high efficiency for modest amounts of extra-dispersion, even when not explicitly accounted for in the parametric model. Pocock et al. (1981) proposed an intermediate solution, via maximum likelihood, to the problem of fitting regression models to tables of frequencies when the residual variation is substantially larger than would be expected from assumptions. Williams (1982) proposed a moment method for logistic linear models, and Breslow (1984) used the method proposed by Pocock et al. (1981) and Williams (1982) for log-linear models. Furthermore, the quasi-likelihood method, which can be considered a moment method, was applied

* Correspondence to: I-BioStat, Universiteit Hasselt, Agoralaan 1, B-3590 Diepenbeek, Belgium. Tel.: +32 11268238.
E-mail address: geert.molenberghs@uhasselt.be (G. Molenberghs).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.06.020

Table 1
Epilepsy data. Number of measurements available at a selection of time points, for both treatment groups separately.

Week	# Observations		
	Placebo	Treatment	Total
1	45	44	89
5	42	42	84
10	41	40	81
15	40	38	78
16	40	37	77
17	18	17	35
20	2	8	10
27	0	3	3

for overdispersion by McCullagh and Nelder (1989) and Wedderburn (1974). The asymptotic properties of all these moment methods for extra-binomial and extra-Poisson variations were studied by Moore (1986).

For modeling longitudinal count data with overdispersion, similarly to Zeger (1988) and Thall and Vail (1990) developed a mixed-effects approach in which the regression coefficients are estimated by generalized estimating equation and the variance component is estimated using method of moments. This may be viewed as an extension of Liang and Zeger's (1986) model for longitudinal count data. Variance components are generally of broad interest (Pryseley et al., 2011).

Besides, Booth et al. (2003) and Molenberghs et al. (2007) brought together both modeling strands and allowed at the same time correlation between repeated measures and overdispersion in the counts. This work was extended by Molenberghs et al. (2010) to data types different from counts. Molenberghs et al. (2007) termed their model the *combined model*. All of these authors conducted parameter estimation and inferences using a likelihood paradigm. In contrast, this paper takes a likelihood perspective. In particular, two versions of a hierarchical Poisson model for longitudinal count data are studied. The first one includes subject-specific random effects to account for subject heterogeneity (a conventional generalized linear mixed model) and the second one includes an additional parameter accounting for overdispersion, generated through an additional gamma distributed random effect (a combined model). The two models are applied to real longitudinal count data and compared using a simulation study.

This paper proceeds as follows. In Section 2, the motivating study is described, which comprises a set of data on epileptic patients. The statistical methodologies is laid out in Section 3. In Section 4, the data set is analyzed, followed by a simulation study in Section 5.

2. The epilepsy data

The data set used in this study is obtained from 89 epileptic patients that are randomized into either placebo or novel anti-epileptic drug (AED), in combination with one or two other AED's after a 12-week run-in period. 45 patients were assigned to the placebo group, the rest to AED. This is a double-blind, parallel group multi-center study. Patients were measured weekly and followed during 16 weeks. That said, some patients were measured up to 27 weeks. The aim of the study was to compare between the groups, the number of seizures experienced during the last week. Note that there are relatively few observations from 20 weeks onwards. Table 1 shows the number of measurements at a selection of time-points. These data were used as one of the three illustrating examples in Booth et al. (2003) who also considered models for longitudinally observed counts that accommodate, at the same time, overdispersion and correlation between repeated measures; for a more elaborate discussion regarding the data, refer to Faught et al. (1996) and Molenberghs et al. (2007). The individual profile curves for both arms is shown in Fig. 1 and reveals substantial variability between subjects; the graphs also show the presence of rather extreme values. It was noticed that there was up and down behavior in the mean evolution. Specifically, on average, there was a substantially higher number of epileptic seizures on week 19 in the placebo group than in the treatment group (Fig. 2). The observed variances at each week are shown in Fig. 2. Notice that there is very high variability in week 19 in the placebo group.

To gain insight into the extent of overdispersion, the sample mean and sample variance at each week for the treatment and placebo group was calculated (Table 2). Clearly, the sample variance is much larger than the sample mean, underscoring the presence of overdispersion in the data. This effect is evident as well from the scale of the mean evolution and variance structure in Fig. 2.

3. A hierarchical Poisson–Normal model with extra-dispersion

Let Y_{ij} represent the number of epileptic seizures that patient i experiences during week j, $i = 1, 2, \ldots, 89$ and $j = 1, 2, \ldots, n_i$, where n_i is the number of repeated measurements for patient i. There are 1419 measurements available in total. We assume the following hierarchical Poisson–Normal model (HPN): $Y_{ij} \sim \text{Poisson}(\lambda_{ij}|b_i)$ with

$$\eta_{ij} = \log(\lambda_{ij}) = \beta_{00} \times I_i + \beta_{01} \times (1 - I_i) + \beta_{10} \times I_i \times t_{ij} + \beta_{11} \times (1 - I_i) \times t_{ij} + b_i, \quad (1)$$