Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 6511–6520
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Likelihood approach for count data in longitudinal experiments

M. Helena Gonçalves[a, b, *], M. Salomé Cabral[c], Maria Carme Ruiz de Villa[d], Eduardo Escrich[e], Montse Solanas[e]

[a] *Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal*
[b] *Departamento de Matemática, FCT, Universidade do Algarve, Campus de Gambelas, 8005-139 Faro, Portugal*
[c] *Centro de Estatística e Aplicações, Departamento de Estatística e Investigação Operacional da Faculdade de Ciências da Universidade de Lisboa, Portugal*
[d] *Facultad de Biologia, Universidad de Barcelona, Spain*
[e] *Facultad de Medicina, Universidad Autónoma de Barcelona, Spain*

Received 16 February 2006; received in revised form 1 March 2007; accepted 1 March 2007
Available online 12 March 2007

Abstract

In many cancer studies and clinical research, repeated observations of response variables are taken over time on each individual in one or more treatment groups. In such cases the repeated observations of each vector response are likely to be correlated and the autocorrelation structure for the repeated data plays a significant role in the estimation of regression parameters. A random intercept model for count data is developed using exact maximum-likelihood estimation via numerical integration. A simulation study is performed to compare the proposed methodology with the traditional generalized linear mixed model (GLMM) approach and with the GLMM when penalized quasi-likelihood method is used to perform maximum-likelihood estimation. The methodology is illustrated by analyzing data sets containing longitudinal measures of number of tumors in an experiment of carcinogenesis to study the influence of lipids in the development of cancer.

Keywords: Longitudinal discrete data; Poisson regression; Repeated measures; Random effects

1. Introduction

Longitudinal count data studies often arise in biostatistical analysis such as those resulting from cancer studies and clinical research. In these studies measurements are collected at several time points on each individual in one or more treatment groups. In such cases the repeated measures of each vector of responses are likely to be correlated and the autocorrelation structure for the repeated data plays a significant role in the estimation of regression parameters. Several approaches have been proposed in the context of count data. Zeger (1988) proposed an estimation equation approach for regression analysis with a time series of counts analogous to the one used by Liang and Zeger (1986). Thall and Vail (1990) considered a mixed effects approach for modeling longitudinal count data with overdispersion, which may be viewed as an extension of the method proposed by Liang and Zeger (1986). Jowaheer and Sutradhar (2002) used generalized estimating equations to model longitudinal count data with overdispersion. Azzalini (1994) proposed an

* Corresponding author. Departamento de Matemática, FCT, Universidade do Algarve, Campus de Gambelas, 8005-139 Faro, Portugal. Tel.: +351 289800905; fax: +351 289800066.
E-mail address: mhgoncal@ualg.pt (M.H. Gonçalves).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.03.002

approach where he made use of the idea of a discrete self-decomposable probability distribution following Steutel and van Harn (1979).

In the present paper, we provide a direct evolution of the model proposed by Azzalini (1994) for Poisson response variables where the serial dependence is assumed to be of Markovian type. We introduce random effects in the linear predictor using exact maximum-likelihood estimation via numerical integration. The maximization procedure for the random effects model involves Gaussian quadrature to calculate the log-likelihood function and the derivatives. One-dimensional integrals are computed based on the `Fortran` subroutine `Dqage` from a `Fortran` subroutine package `QUADPACK`. The maximization procedure is performed in R; the observed Fisher information matrix is computed via numerical differentiation of the first derivatives. A brief simulation study is carried out to compare our methodology to the traditional approach (generalized linear mixed model, GLMM), which ignores the conditional independence between repeated measures in terms of numerical analysis, as well as to the GLMM when the penalized quasi-likelihood (PQL) method is used to perform maximum-likelihood estimation. For GLMM approach the estimates were obtained through the function glmmML in the R package glmmML. The function glmmPQL in the R MASS library is used to obtain the PQL estimates (fitted by maximum likelihood). The proposed methodology is illustrated by analyzing data sets containing longitudinal measures of number of tumors in an experiment of carcinogenesis to study the influence of lipids in the development of cancer. The paper is organized as follows: Section 2 gives a summary of the model proposed and the derivation of the methodology. Section 3 reports a simulation study. Section 4 illustrates the method with the analysis of real data. Section 5 concludes the paper.

2. Random intercept model for longitudinal count data

2.1. Introduction

To establish notation, denote by $y_{it}(t = 1, \ldots, T_i)$ the response value at time t from subject $i\,(i = 1, \ldots, n)$, and by Y_{it} its generating random variable which has a Poisson distribution whose mean value is $E(Y_{it}) = \theta_{it}$. We shall refer collectively to the sequence $(y_{i1}, \ldots, y_{iT_i})$ as the ith individual profile, possibly with some missing data. Associated with each observation time and each subject, a set of p covariates is available, denoted by x_{it} and β as the p-vector of unknown parameters. Our aim is to introduce a Poisson regression which links the covariates and the probability distribution of the response, in the form

$$\ln\{E(Y_{it})\} = \ln(\theta_{it}) = x_{it}^{\top}\beta, \tag{1}$$

allowing also some form of dependence among observations of the same individual. Although the logarithmic link is the natural choice of count data we can also use the identity link.

2.2. A model for longitudinal count data

We consider the model proposed by Azzalini (1994) where the serial dependence is assumed to be of Markovian type. Since our work builds on a direct evolution of the model proposed on that paper, it makes our exposition simpler to start by summarizing the scheme adopted there. To simplify notation, we drop temporarily the subscript i, since individuals are assumed to behave independently from each other and the same stochastic model can then simply be replicated n times. Following Steutel and van Harn (1979), if $\rho \in (0, 1)$ and W is a random variable taking values of the non-negative integers, $\rho \circ W$ is defined as

$$\rho \circ W = \sum_{h=1}^{W} Z_h, \tag{2}$$

where $Z_1, Z_2, \ldots$ is a sequence of independent Bernoulli variables with common probability of success ρ, $\Pr(Z_h = 1) = 1 - \Pr(Z_h = 0) = \rho$. It then follows that $\rho \circ W \in \mathbb{N}_0$ with $1 \circ W = W$, $0 \circ W = 0$ and $E(\rho \circ W) = \rho(EW)$ as in scalar multiplication.

Suppose that $\theta_1, \theta_2, \ldots$ is a sequence of given positive real numbers. Consider a probability model of the form:

$$Y_t = \rho \circ Y_{t-1} + \varepsilon_t \quad (t = 2, 3, \ldots, T), \tag{3}$$