Available online at www.sciencedirect.com
ScienceDirect
Contemporary Clinical Trials 29 (2008) 547–554
Contemporary Clinical Trials
www.elsevier.com/locate/conclintrial

Modelling overdispersion in longitudinal count data in clinical trials with application to epileptic data

Ali Reza Fotouhi *

University College of the Fraser Valley, 33844 King Road, Abbotsford, Canada BC, V2S 7M8

Received 23 March 2007; accepted 19 January 2008

Abstract

A family of multi-level models with different types of random error components in the linear predictor is presented for analysing longitudinal count data in clinical trials. These models account for overdispersion, heterogeneity, serial correlation, and heteroscedasticity. The proposed models are applied to epileptic seizure count data and illustrated in a simulation study. The effects of omitted variables, link function, outliers, and initial conditions on overdispersion are investigated. It has been shown that proper introduction of the error component in the linear predictor overcomes the problem of overdispersion arising from the omitted variables. We use three model checking criteria deviance, variance inflation factor, and global goodness-of-fit tests based on Bayesian probability to identify the best structure of the error term in the linear predictor. Further, Markov Chain Monte Carlo method using Gibbs sampling is used as estimation approach.

Keywords: Overdispersion; Random effects; Serial correlation; Measurement error; Markov Chain Monte Carlo; Variance inflation factor; Initial conditions; Bayesian probability; Deviance; Simulation

1. Introduction

In applying standard generalized linear models it is often found that the data exhibit greater variability than is predicted by the implicit mean–variance relationship. This phenomenon of overdispersion has been widely considered in the literature, particularly in relation to the Poisson distribution. Overdispersion arises commonly through the omission from the regression models of important explanatory variables, existence of outliers, and using inappropriate link function. This violation of the standard model assumptions inflates the residual deviance, produces large residuals, underestimates standard deviations of the estimated parameters, and may lead to biases in the estimates themselves if the omitted variables are correlated with those in the model. The most important consequence of overdispersion is overstating the significance of explanatory variables.

The Poisson distribution is usually assumed as the distribution of counts in analysing longitudinal count data. In clinical trials occurring of events are usually counted over time for different individuals in the control and the treatment groups. If the data show greater variability than is predicted by a Poisson distribution then this extra variability, overdispersion, should be considered before the Poisson model is fitted to the data. The epileptic seizure count data, Thall and Vail [1], is an

* Fax: +1 604 855 7614.
E-mail address: ali.fotouhi@ucfv.ca.

1551-7144/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2008.01.005

example which exhibits a high degree of overdispersion. This extra Poisson variation should be considered in the proposed model to access the proper interpretation of the treatment effect.

In order to analyse overdispersed data we can broadly categories the approaches into two groups. The first group assumes some more general form for the variance function with additional parameters and use quasi-likelihood approach was first introduced by Wedderburn [2]. Liang and Zeger [3] have proposed quasi-likelihood models that describe the correlation structure among the responses, while also taking overdispersion into account. Thall and Vail [1] have introduced some covariance models for analysing longitudinal count data with overdispersion.

In cross sectional studies, if the extra variation is not related to the explanatory variables and the variance is proportional to the mean, then the quasi-likelihood estimate of the parameter (McCullagh and Nelder [4]) is the same as the Maximum Likelihood Estimate (MLE) from a Poisson model. The standard deviation of the quasi-likelihood estimate may be obtained by multiplying those from the Poisson model by an estimate of the dispersion parameter, a simple scale factor obtained by dividing the residual deviance or the Pearson χ^2 by the residual degrees of freedom.

The second group assumes a multi-level model for the response variable with the model parameter itself having some distribution. In both cross sectional and longitudinal data analysis, if the extra variation among the counts is related to the explanatory variables then the only satisfactory course of action is to attempt to identify the missing variables which explain this variation. Introducing an appropriate random variable as an error term in the linear predictor may model this extra variation. Random effects and serial correlation models are special cases of this approach. The Random effects model has been widely used to control for heterogeneity and the omitted explanatory variables. Recently, Jowaheer and Sutradhar [5] have used this approach and have introduced a negative binomial model for analysing longitudinal count data with overdispersion.

For multi-level models and in simple cases full MLE may be possible. For example, for the Poisson-Gamma model, which is widely used for analysing longitudinal count data, the distribution of the mixing distribution is conjugate to the distribution of the response variable, and then the MLE approach is straight forward. However, Jowaheer and Sutradhar [5] have used the generalized estimating equation approach for their analysis. The approximation methods are often used when mixing distribution is not conjugate to the response distribution. An example is a Poisson-Normal model in which the normal distribution is not conjugate to the Poisson distribution. Fotouhi [6] employed this model in investigating the initial conditions problem in analysing longitudinal count data. The non-parametric approach can also be used when no specific distribution is assumed for the error term. Alfo and Aitkin [7] have applied this method for analysing the longitudinal count data with baseline information.

In this paper we introduce the sources of overdispersion in hierarchical models of longitudinal count data. We show that the correct structure of the error term in the linear predictor is essential to overcome the problem of overdispersion. To illustrate the performance of the proposed models an application of the analysis of epileptic seizure count data, Thall and Vail [1], arising in a study of progabide as an adjuvant anti-epileptic chemotherapy is presented. We show that the problem of overdispersion, which leads to over-estimating the treatment effect, can be solved by introducing an appropriate error term in the linear predictor. The proposed methods are also investigated in a simulation study to see if the results from an application to real data are consistent in a realistic situation. The Markov Chain Monte Carlo (MCMC) method is used to estimate the parameters. Fotouhi [8] has shown that this approach performs very well in fitting multi-level models especially for analysing longitudinal data.

In Section 2 we present the epileptic data which attempts to show the existence of overdispersion, heteroscedasticity, and within-patient dependence. In Section 3 we discuss the theory, estimation procedure, and the model checking criteria. In Section 4 we analyse the epileptic data and compare it with the other results. Section 5 reports the simulation studies. A concluding discussion is given in Section 6.

2. Data

In this paper we analyse the data arising from a clinical trial of 59 epileptics reported in Table 1. Patients suffering from simple or complex partial seizures were randomized to receive either the anti-epileptic drug progabide or a placebo, as an adjuvant to the standard chemotherapy. At each of four successive post-randomization clinic visits, the number of seizures occurring over the pervious two weeks was reported. Although each patient subsequently was crossed over to the other treatment, we shall consider only the four