Statistical Methodology 30 (2016) 1–14

Contents lists available at ScienceDirect

Statistical Methodology

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

Multilevel zero-inflated Generalized Poisson regression modeling for dispersed correlated count data

Afshin Almasi [a], Mohammad Reza Eshraghian [b,*], Abbas Moghimbeigi [c], Abbas Rahimi [b], Kazem Mohammad [b], Sadegh Fallahigilan [d]

[a] *Department of Biostatistics and Epidemiology, School of Public Health, Kermanshah University of Medical Sciences, Kermanshah, Iran*

[b] *Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran*

[c] *Modeling of Non-communicable Disease Research Center, Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*

[d] *Department of Statistics, Razi University, Kermanshah, Iran*

ARTICLE INFO

Article history:
Received 23 February 2015
Received in revised form
26 October 2015
Accepted 5 November 2015
Available online 14 November 2015

Keywords:
Count data
EM algorithm
Multilevel
Generalized Poisson regression
Zero-inflation
Dispersion
Monte Carlo simulation

ABSTRACT

Poisson or zero-inflated Poisson models often fail to fit count data either because of over- or underdispersion relative to the Poisson distribution. Moreover, data may be correlated due to the hierarchical study design or the data collection methods. In this study, we propose a multilevel zero-inflated generalized Poisson regression model that can address both over- and underdispersed count data. Random effects are assumed to be independent and normally distributed. The method of parameter estimation is EM algorithm base on expectation and maximization which falls into the general framework of maximum-likelihood estimations. The performance of the approach was illustrated by data regarding an index of tooth caries on 9-year-old children. Using various dispersion parameters, through Monte Carlo simulations, the multilevel ZIGP yielded more accurate parameter estimates, especially for underdispersed data.

* Corresponding author. Tel.: +98 9121078670.
E-mail address: eshraghianmr@yahoo.com (M.R. Eshraghian).

http://dx.doi.org/10.1016/j.stamet.2015.11.001
1572-3127/

1. Introduction

Statistical models that address count data have been implemented in many areas such as the insurance industry, risk classification, health care, accident frequencies, dental epidemiology, medicine, etc. One of the most frequently used of these types of models is Poisson regression. The major problems hindering a valid fit in this model type tend to be the presence of excess zeros and over- or underdispersion with respect to a Poisson distribution. A common approach for overcoming the problem of excess zeros is the utilization of zero-inflated (ZI) regression models such as the zero-inflated Poisson (ZIP) [15]. Inappropriate use of a Poisson or ZIP regression model, however, may cause inaccurate parameter estimations. Various distributions have been proposed to handle overdispersion in count data, including weighted Poisson [24,13], Quasi-Poisson [29], Poisson Inverse Gaussian [32], Poisson lognormal [2], Poisson–Lindley [7], Negative Binomial-Lindley [37], Negative Binomial (NB) [10] and Generalized Poisson (GP) [31]. In addition to a zero-inflated negative binomial (ZINB) regression model, which can be used in instances of excess zeros and overdispersion in data, the increasingly popular zero-inflated generalized Poisson (ZIGP) models can also be applied to both over- and underdispersed count data [31,6].

Comparisons between GP and NB reveal slight differences in many aspects. For instance, the GP distribution has a heavier tail, while the NB distribution has a larger mass at zero. However, their zero-inflated distributions, with masses at zero and fixed means and variances, can differ even more from each other [11]. In addition, there is a relatively unknown number of situations in which the iterative estimation technique for estimating parameters of a ZINB regression model may fail to converge [6] while the ZIGP model may converge more often. Prior studies have described the classical Generalized Poisson regression model, named GP-1 and introduced by Consul and Jain [5], and its properties [4,27]. A different parameterization of a Generalized Poisson model is called the GP-2 regression model [31,34]. It is interesting to note that the GP-1 and the GP-2 regression models are natural extensions of the Poisson regression model. Recently, the functional form of the ZIGP regression model was extended and introduced as the ZIGP-P regression model, where the ZIGP-1 and ZIGP-2 models are special cases of the ZIGP-P model when $P = 1$ and $P = 2$, respectively [39]. It is important to note that the GP-1 and GP-2 distributions discussed in this paper are different from the Compounded Geometric (CG) and G(p) distributions considered in Chowdhury et al. [3].

In models used for count data, the assessment of over- or underdispersion and zero-inflation is very important. Gupta and Gupta et al. [8] and Famoye and Singh [6] developed score tests to assess both zero-inflation and dispersion in a ZIGP regression model. Yang and Hardin et al. [35] proposed a score test for overdispersion based on the GP-1 and GP-2 models; the power of this proposed test is higher than that for the likelihood ratio (LR) and Wald test and is more appropriate in practice [34]. Xie and Wei et al. [33] extended several score tests for mixed regression models, i.e., the ZIP versus the ZIGP regression model. Zamani and Ismail [38] proposed another score test applicable for the ZIP regression model against the ZIGP and proved that this test was equal to the ZIP regression model against the ZINB. Another approach called the hurdle model, like the ZI model, is a 2-part count regression method that addresses the phenomenon of excess zeros in data. However, hurdle models are different from the ZI models. The first component of a hurdle model, typically logistic regression, addresses the probability of a zero count (as opposed to an 'excess zero') so that it depends on the prevalence (or incidence) in the overall population because it targets all zero counts. The second part of a hurdle model is used for the mean count among subjects with non-zero counts [25].

Multilevel or random effect models are key tools for research designs where participant data are organized at more than one level (i.e., nested data). Such data arise routinely in various fields, for instance, in medical research with patients nested within physicians or hospitals and biological research involving analysis of dental problems with teeth nested within different mouths. Often, because of the use of hierarchical designs or differences in data collection procedures such as a longitudinal study design in which data are gathered for the same subjects repeatedly over a period of time, the collected data may show zero-inflation and lack of independence as a result of its inherent structural correlation and/or underlying heterogeneity. Lee and Wang et al. [17] proposed a class of multilevel ZIP (MZIP) models for such data. In addition, Moghimbeigi and Eshraghian et al. [23] introduced a multilevel ZINB (MZINB) regression model for overdispersed count data. Instead of