Computational Statistics and Data Analysis 71 (2014) 151–158

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Zero-inflated Poisson regression mixture model

Hwa Kyung Lim *, Wai Keung Li, Philip L.H. Yu

Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 5 July 2012
Received in revised form 23 June 2013
Accepted 23 June 2013
Available online 29 June 2013

Keywords:
Zero-inflation
Heterogeneity
Finite mixture model
Poisson
EM algorithm

ABSTRACT

Excess zeros and overdispersion are common phenomena that limit the use of traditional Poisson regression models for modeling count data. Both excess zeros and overdispersion caused by unobserved heterogeneity are accounted for by the proposed zero-inflated Poisson (ZIP) regression mixture model. To estimate the parameters of the model, an EM algorithm with an embedded iteratively reweighted least squares method is implemented. The parameter estimation performance of the proposed model is evaluated through simulation studies. The ZIP regression mixture model is applied to the DMFT index dataset, which contains excess zeros and overdispersion. Comparisons of several other models commonly used for such data with the ZIP regression mixture model show that, in general, the latter model fits the data well.

1. Introduction

Modeling count data is a topic of major interest in fields such as sociology, engineering, medical studies and others. The classical Poisson regression model for count data is often of limited use in these disciplines because empirical count data typically exhibit overdispersion (i.e., the variance of the response variable exceeds the mean). This phenomenon often results from unobserved heterogeneity, which occurs when the sample of responses are drawn from a population consisting of several sub-populations. Mixtures of Poisson distributions have been widely used to deal with this problem. For example, a finite Poisson mixture model with K components explains the population by giving weights π_k to sub-populations with means λ_k, $k = 1, \ldots, K$. This approach also provides a natural framework to classify observations into the components of the mixture model. Poisson mixtures were first proposed by Simar (1976) and Laird (1978). Finite mixtures of Poisson regression models with constant weight parameters have been developed by Wedel et al. (1993), Brännaäs and Rosenqvist (1994), Wang et al. (1996), and Alfò and Trovato (2004). Wang et al. (1998) discuss finite mixed Poisson regression models that incorporate covariates in the weight parameters. As an alternative to handling overdispersion, a negative binomial (NB) regression model can be used since it allows the variance to be larger than the mean.

The count variable of interest may contain more zeros than expected under a Poisson model, which is commonly observed in many applications. For instance, the DMFT index, analyzed in Section 5, indicates the number of defective teeth in adolescents. As expected, a large number of subjects have no defective teeth, which illustrates an occurrence of zero-inflation. A popular approach to modeling excess zeros is to use a zero-inflated Poisson (ZIP) regression model, as discussed by Lambert (1992). The ZIP distribution is a mixture of a Poisson distribution and a degenerate distribution at zero. This regression setting allows for covariates in both the Poisson mean and weight parameter. Böhning (1998) and Ridout et al. (1998) provide reviews of the related literature and present examples from a wide variety of disciplines.

Furthermore, if overdispersion remains even after modeling excess zeros, a zero-inflated negative binomial (ZINB) regression model can provide a good solution. However, if a population has excess zeros and several sub-populations in non-zero counts, a single component of the ZINB regression model may not be sufficient to describe the non-zero counts. In this paper, we propose the ZIP regression mixture model for heterogeneous count data with excess zeros.

* Corresponding author.
E-mail addresses: hklim@korea.ac.kr (H.K. Lim), hrntlwk@hku.hk (W.K. Li), plhyu@hku.hk (P.L.H. Yu).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.06.021

The paper is organized as follows. We describe the ZIP regression mixture model in Section 2. The EM algorithm for model fitting is described in Section 3. Several simulation studies assessing the performance and sensitivity of parameter estimation are presented in Section 4, and Section 5 demonstrates real data applications of the model. Finally, we conclude with a discussion in Section 6.

2. ZIP regression mixture model

Suppose that a count response variable Y follows a ZIP mixture distribution:

$$P(Y=y)=\begin{cases}\pi_1+\pi_2 e^{-\lambda_2}+\cdots+\pi_K e^{-\lambda_K}, & y=0\\ \pi_2\dfrac{e^{-\lambda_2}\lambda_2^y}{y!}+\cdots+\pi_K\dfrac{e^{-\lambda_K}\lambda_K^y}{y!}, & y>0\end{cases} \tag{1}$$

where K is the number of mixing components, λ_k is the mean, and π_k is the mixing weight of component k such that $0<\pi_k<1$, $k=1,\ldots,K$, and $\sum_{k=1}^{K}\pi_k=1$. The weight π_1 determines the proportion of excess zeros compared with an ordinary Poisson mixture model. If K is equal to two, the ZIP mixture distribution in Eq. (1) is reduced to the ZIP distribution (Lambert, 1992).

To allow the mean and the mixing weight to depend on covariates, we model $\{\lambda_k\}_{k=2}^K$ and $\{\pi_k\}_{k=1}^K$ using the following regression models that assume $\log(\lambda_k)$ and the multinomial logit of π_k to be linear functions of covariates:

$$\log(\lambda_{ik})=\mathbf{x}_i\beta_k,\quad i=1,\ldots,N,\ k=2,\ldots,K \tag{2}$$

$$\pi_{ik}(\mathbf{w}_i,\gamma)=\frac{\exp(\mathbf{w}_i\gamma_k)}{1+\sum_{k=2}^{K}\exp(\mathbf{w}_i\gamma_k)},\qquad \pi_{i1}(\mathbf{w}_i,\gamma)=1-\sum_{k=2}^{K}\pi_{ik}(\mathbf{w}_i,\gamma), \tag{3}$$

where $\mathbf{x}_i=(x_{i1},\ldots,x_{ip})$ and $\mathbf{w}_i=(w_{i1},\ldots,w_{iq})$ are $1\times p$ and $1\times q$ row vectors of covariates (including an intercept), respectively, and β_k and γ_k are the corresponding $p\times 1$ and $q\times 1$ vectors of regression coefficients for the kth component, respectively. Note that the mixing probability of the first component $\pi_{i1}(\mathbf{w}_i,\gamma)$ is the probability of excess zeros, and is taken as the baseline for the multinomial logit. That is, the logit for the other components relative to π_{i1} is $\log(\pi_{ik}/\pi_{i1})=\mathbf{w}_i\gamma_k$, $k=2,\ldots,K$.

The generalized ZIP (GZIP) regression mixture model can be formulated as follows:

$$P(Y=y_i)=\pi_{i1}(\mathbf{w}_i,\gamma)I_{(y_i=0)}+\sum_{k=2}^{K}\pi_{ik}(\mathbf{w}_i,\gamma)Pois(y_i\mid\lambda_{ik}(\mathbf{x}_i,\beta_k)), \tag{4}$$

where $I_{(\cdot)}$ is 1 if the specified condition is satisfied and 0 otherwise, and $Pois(y_i\mid\lambda_{ik}(\mathbf{x}_i,\beta_k))$ denotes the Poisson probability mass function of y_i with mean $\lambda_{ik}(\mathbf{x}_i,\beta_k)$. A special case of the above model will be obtained if the mixing weights $\pi_{ik}(\mathbf{w}_i,\gamma)$ are assumed to be constant functions of the covariates, $\mathbf{w}_i$. In that case, the ZIP with fixed weights (FZIP) regression mixture model can be formulated as follows:

$$P(Y=y_i)=\pi_1 I_{(y_i=0)}+\sum_{k=2}^{K}\pi_k Pois(y_i\mid\lambda_{ik}(\mathbf{x}_i,\beta_k)). \tag{5}$$

If both π_{ik} and λ_{ik} are constant functions, the GZIP mixture model reduces to the standard Poisson mixture model, denoted by

$$P(Y=y_i)=\sum_{k=1}^{K}\pi_k Pois(y_i\mid\lambda_k). \tag{6}$$

Note that, the first component (a degenerate distribution with all mass π_1 at $y_i=0$) in Eq. (4) can be regarded as a Poisson distribution with a mean of $\lambda_1=0$, because $Pois(y_i=0\mid\lambda_1=0)=1$ and $Pois(y_i\neq 0\mid\lambda_1=0)=0$.

In the following section, we describe an estimation method based on the EM algorithm for the GZIP regression mixture model given by Eq. (4).

3. Model estimation

The EM algorithm can be applied to obtain the maximum likelihood estimates (MLEs) in a finite mixture model of arbitrary distributions (McLachlan and Krishnan, 1997). Let the number of components, K, be fixed and known, and $\mathbf{z}_i=(z_{i1},\ldots,z_{iK})$ be the latent vector of component indicator variables, where

$$z_{ik}=\begin{cases}1, & i\text{th subject comes from the latent } k\text{th component}\\ 0, & \text{otherwise.}\end{cases}$$