Computational Statistics and Data Analysis 55 (2011) 765–773

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Model selection for zero-inflated regression with missing covariates

Xue-Dong Chen [a,*], Ying-Zi Fu [b]

[a] School of Science, Huzhou Teachers College, No. 1 Xueshi Road, Huzhou, Zhejiang Province, 313000, China
[b] School of Science, Kunming Science and Technology University, Yunnan 650093, China

ARTICLE INFO

Article history:
Received 2 May 2009
Received in revised form 11 June 2010
Accepted 25 June 2010
Available online 5 July 2010

Keywords:
Zero-inflation
Missing data
Model selection
AIC
EM algorithm

ABSTRACT

Count data are widely existed in the fields of medical trials, public health, surveys and environmental studies. In analyzing count data, it is important to find out whether the zero-inflation exists or not and how to select the most suitable model. However, the classic AIC criterion for model selection is invalid when the observations are missing. In this paper, we develop a new model selection criterion in line with AIC for the zero-inflated regression models with missing covariates. This method is a modified version of Monte Carlo EM algorithm which is based on the data augmentation scheme. One of the main attractions of this new method is that it is applicable for comparison of candidate models regardless of whether there are missing data or not. What is more, it is very simple to compute as it is just a by-product of Monte Carlo EM algorithm when the estimations of parameters are obtained. A simulation study and a real example are used to illustrate the proposed methodologies.

1. Introduction

Count data with many zeros are commonly encountered in the fields of medical trials, public health, surveys, and environmental studies. A lot of models have been found effective for the analysis of such data. For example, see Lambert (1992), Famoye and Singh (2006), Xiang et al. (2007), and Czado et al. (2007). The utilization of the above regression models is based on the fact that they can handle zero-inflation or overdispersion, which the count data very often exhibit. In Bayesian framework, Angers and Biswas (2003), and Ghosh et al. (2006) developed some Bayesian estimation methods for different zero-inflated regression models respectively. In count data analysis, there are two important issues, one is whether there exist extra zeros or not which result in zero-inflation in the data and the other is how to select the best model among so many candidate models. To answer these questions, several approaches such as score test, likelihood ratio test, Bayesian test and model comparison using AIC criterion have been proposed by Famoye and Singh (2006), Xiang et al. (2007), Bhattacharya et al. (2008), Czado et al. (2007) and so on.

In addition, missing data often arise in various settings, including surveys, clinical trials, and environmental research. It is well known that the analyses based on deleting the cases with missing values, usually referred to as a complete case analysis, will inevitably leads to the bias and inefficient estimates of regression coefficients and their standard errors. As a result, some methods are developed for taking advantage of the incomplete information involved in the missing data instead of ignoring them. The literature on this issue is extensive, such as Ibrahim et al. (1999), Ibrahim et al. (2001), Huang et al. (2005), Chen and Ibrahim (2006) and so on. An overview of common approaches for inference in GLMs with missing covariates can be found in Little and Rubin (2002) and Ibrahim et al. (2005) respectively.

Although there have been so many theories and methods cited before that can be used to deal with the above two issues separately, there is no work done on zero-inflated regression models with missing covariates. Furthermore, the existing

* Corresponding author. Tel.: +86 572 2365697.
E-mail address: xdchen@hutc.zj.cn (X.-D. Chen).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.06.023

maximum likelihood estimation methods are not applicable for zero-inflated regression models with missing covariates, not to mention the problem of model selection for this case. It is necessary for us to establish a new approach to cope with more complex models. The main purpose of this paper is to develop a novel model selection criterion for zero-inflated regression models with missing covariates on the basis of Claeskens and Consentino (2008).

Inspired by the idea of data augmentation, we propose a modified version of Monte Carlo EM algorithm and the ML estimation of model parameters are obtained, then we develop a new model selection criterion in line with AIC on the basis of Q-function, which is based on the E-step of the MCEM algorithm rather than the more complicated observed data log-likelihood. Our method is rather effective for comparison among several candidate models whether the zero-inflation exists or not as well as whether there are missing observations or not, moreover, it is very easy to evaluate as it is just a by-product of Monte Carlo EM algorithm when the MCEM algorithm is converged.

The rest of this paper is organized as follows. In Section 2, we introduce the models of zero-inflated power series (ZIPS) with missing covariates. In Section 3, Based on data augmentation, the E-step and M-step of Monte Carlo EM algorithm is derived and the Q-function is obtained. A new model selection criteria for ZIPS regression model with missing covariates is described in Section 4. To illustrate the proposed methodologies, a simulation study and a real example of surveys data are presented in Section 5.

2. Models and notation

2.1. Regression models for ZIPS distribution

In general, zero-inflated data can be viewed as a mixture of a degenerate distribution with mass at zero and a non-degenerate distribution such as the binomial or Poisson distribution and it can be given as follows

$$p(y|\omega,\theta)=\begin{cases}\omega+(1-\omega)f(0|\theta), & y=0,\\ (1-\omega)f(y|\theta), & y>0,\end{cases} \tag{1}$$

where $f(\cdot|\theta)$ denotes the non-degenerate distribution, $\theta\in\Theta$, the parameter space of $f(\cdot|\theta)$, and the mixing parameter ω ranges over the interval $-f(0|\theta)/(1-f(0|\theta))<\omega<1$. Generally speaking, any discrete distribution could be used for $f(\cdot|\theta)$. For simplicity and some theoretical consideration, we focus on a flexible class of discrete distribution, namely, the power series (PS) distribution given by

$$f(y|\theta)=\frac{a(y)\theta^y}{g(\theta)}, \quad y=0,1,2,\ldots, \tag{2}$$

in which $a(y)$ is a known function and can be viewed as the coefficient of the power series with respect to θ, and $g(\theta)=\sum_{y=0}^{\infty}a(y)\theta^y$ is the normalizing constant. Obviously, Poisson distribution $\mathrm{P}(\theta)$ and negative binomial distribution $\mathrm{NB}(\theta,\sigma)$ belong to $\mathrm{PS}(\theta)$ distribution with $a(y)=\frac{1}{y!}$, $g(\theta)=e^{\lambda}$ and $a(y)=\frac{\Gamma(\sigma+y)}{\Gamma(y+1)}$, $g(\theta)=\Gamma(\sigma)(1-\theta)^{-\sigma}$ respectively, where $\sigma>0$ is a dispersion parameter and, for simplicity, it is assumed to be known or can be estimated separately in the regression. Then the ZIPS distribution defined by (1) and (2) can be represented as $Y\sim\mathrm{ZIPS}(\omega,\theta)$, note that $E(Y)=(1-\omega)\mu(\theta)$, where $\mu(\theta)=\theta g'(\theta)/g(\theta)$ denote the means of regular $\mathrm{PS}(\theta)$ distribution. It is apparent that ZIPS includes a lot of common distributions such as zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) with $f(y|\theta)$ follows Poisson distribution $\mathrm{P}(\theta)$ and negative binomial distribution $\mathrm{NB}(\theta,\sigma)$ respectively.

For independently distributed responses $\{Y_i,\ i=1,\ldots,n\}$ sampled from $\mathrm{ZIPS}(\omega_i,\theta_i)$, the ZIPS regression model is defined as

$$\log[\mu(\theta_i)]=\boldsymbol{x}_i^T\boldsymbol{\beta}\quad\text{and}\quad\log\left(\frac{\omega_i}{1-\omega_i}\right)=\boldsymbol{z}_i^T\boldsymbol{\gamma} \tag{3}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are covariate vectors related to the mean $\mu(\theta_i)$ and zero proportion parameters ω_i respectively, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the corresponding vector of regression coefficients. For simplicity of notation, we assume that covariate matrix $\boldsymbol{X}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^T$ and $\boldsymbol{Z}=(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n)^T$ are the same throughout, and it can be easily extended to the case in which $\boldsymbol{X}$ and $\boldsymbol{Z}$ are different. Furthermore, we assume that $\boldsymbol{x}_i=(x_{i1},\ldots,x_{iq})$, $i=1,\ldots,n$ are partially missing, so the design matrix $\boldsymbol{X}$ can be split into two parts, $\boldsymbol{X}=(\boldsymbol{X}_{obs},\boldsymbol{X}_{mis})$ and the i-th rows of $\boldsymbol{X}_{obs}$ and $\boldsymbol{X}_{mis}$ are denoted by, respectively, $\boldsymbol{x}_{obs,i}$ and $\boldsymbol{x}_{mis,i}$. From (1)–(3), we write

$$p(y_i|\omega_i,\theta_i)\equiv p(y_i|\boldsymbol{x}_i,\boldsymbol{\beta},\boldsymbol{\gamma}), \tag{4}$$

and call the model described in (4) ZIPS regression model. In order to obtain unbiased and efficient estimates of the coefficients in this model, we need to specify a parametric model for the covariates with missing data.