Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Generalized endpoint-inflated binomial model

Q1 Guo-Liang Tian [a], Huijuan Ma [b,*], Yong Zhou [c,d], Dianliang Deng [e]

[a] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, PR China
[b] Department of Statistics and Finance, The University of Science and Technology of China, Hefei, Anhui, PR China
[c] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, PR China
[d] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, PR China
[e] Department of Mathematics and Statistics, University of Regina, Saskatchewan, Canada

ARTICLE INFO

Article history:
Received 16 May 2014
Received in revised form 8 March 2015
Accepted 9 March 2015
Available online xxxx

Keywords:
Endpoint-inflated binomial distribution
Expectation–maximization algorithm
Multinomial logistic regression model
Stochastic representation
Zero-inflated binomial distribution

ABSTRACT

To model binomial data with large frequencies of both zeros and right-endpoints, Deng and Zhang (in press) recently extended the zero-inflated binomial distribution to an *endpoint-inflated binomial* (EIB) distribution. Although they proposed the EIB mixed regression model, the major goal of Deng and Zhang (2015) is just to develop score tests for testing whether endpoint-inflation exists. However, the distributional properties of the EIB have not been explored, and other statistical inference methods for parameters of interest were not developed. In this paper, we first construct six different but equivalent stochastic representations for the EIB random variable and then extensively study the important distributional properties. Maximum likelihood estimates of parameters are obtained by both the Fisher scoring and expectation–maximization algorithms in the model without covariates. Bootstrap confidence intervals of parameters are also provided. Generalized and fixed EIB regression models are proposed and the corresponding computational procedures are introduced. A real data set is analyzed and simulations are conducted to evaluate the performance of the proposed methods. All technical details are put in a supplemental document (see Appendix A).

1. Introduction

Modeling count data with many zeros are common in many fields including medicine, public health, epidemiology, ecology, sociology, psychology, econometrics, agriculture, engineering, manufacturing, and road safety. A large number of statistical methodologies have been developed to analyze such data. Existing literature on this issue can be roughly categorized into two classes. The first class focuses on the development of distributional properties and the relevant statistical inference methods without considering the covariate. The other class is to propose various zero-inflated regression models to account for the covariate effect. The *zero-inflated Poisson* (ZIP) regression model (Mullahy, 1986; Lambert, 1992) and its variants are quite popular in practice. When the counts have an upper bound, the ZIP regression model is no longer appropriate. Hall (2000) and Vieira et al. (2000) introduced a *zero-inflated binomial* (ZIB) regression model (also called Bernoulli–Binomial mixture model) while Hall and Berenhaut (2002) developed ZIB mixed models. Ospina and Ferrari (2010) proposed zero- or/and one-inflated beta distributions while Ospina and Ferrari (2012) studied a general class of regression models for continuous proportions when the data contain many zeros or ones. Adell et al. (2012) proposed a kind of one-inflated bivariate beta distribution to analyze matching scores related to retinal image identification in lambs.

* Corresponding author.
E-mail address: mhj@mail.ustc.edu.cn (H. Ma).

http://dx.doi.org/10.1016/j.csda.2015.03.009
0167-9473/

Please cite this article in press as: Tian, G.-L., et al., Generalized endpoint-inflated binomial model. Computational Statistics and Data Analysis (2015), http://dx.doi.org/10.1016/j.csda.2015.03.009

However, in practice, we may encounter discrete proportion data with extra zeros and extra ones (or extra right endpoints). For example, in the epidemiological study, the incidence of an infective disease in some families is either zero or 100% during a period of infection (Deng and Zhang, in press). That is, besides the structural zeros and the structural right-endpoints, there are not only extra zeros (i.e., left-endpoints) but also extra right-endpoints. As the second example, in Section 6 we shall introduce the whitefly data set, in which the number of surviving whiteflies demonstrate both extra zeros and extra right-endpoints because of the efficacy of the pesticide. In other words, if the pesticide has a strong effect, it usually kills all the whiteflies in one cage, causing more zeros in the real data; while in the control group where no pesticide has been administrated, all the adult whiteflies will survive, resulting in more right-endpoints. In fact, there are 640 observations with 339 zeros (53%) and 76 right-endpoints (12%).

Thus, it is inappropriate to model such binomial data with excess of zeros and right-endpoints by using ZIB distribution and zero- or/and one-inflated beta distribution. As a generalization of the widely discussed ZIB, a so-called *zero–one inflated binomial* (ZOIB) distribution was proposed recently by Deng and Zhang (in press), which was the unique paper involving the ZOIB model to date. To avoid confusion, hereafter, we call the ZOIB distribution the *endpoint-inflated binomial* (EIB) distribution. Although they proposed the EIB mixed regression, the major goal of Deng and Zhang (in press) is just to develop score statistics for testing whether endpoint-inflation exists. However, the distributional theory and corresponding properties of the EIB have not yet been explored, and other statistical inference methods for parameters of interest were not well developed. The main objective of this paper is to fill the gap.

For convenience, in this paper we denote a random variable ξ following a degenerate distribution with all mass at a single point c by $\xi \sim \text{Degenerate}(c)$, whose *probability mass function* (pmf) is $\Pr(\xi = c) = 1$. Let $\xi_0 \sim \text{Degenerate}(0)$, $\xi_1 \sim \text{Degenerate}(m)$, $X \sim \text{Binomial}(m, p)$ and they are independent. A discrete random variable Y is said to have an EIB distribution, denoted by $Y \sim \text{EIB}(\phi_0, \phi_1; m, p)$, if its pmf is (Deng and Zhang, in press)

$$\begin{aligned} f(y|\phi_0, \phi_1; m, p) &= \phi_0 \Pr(\xi_0 = y) + \phi_1 \Pr(\xi_1 = y) + \phi_2 \Pr(X = y) \\ &= \begin{cases} \phi_0 + \phi_2(1-p)^m, & \text{if } y = 0, \\ \phi_2 \binom{m}{y} p^y (1-p)^{m-y}, & \text{if } y = 1, \ldots, m-1, \\ \phi_1 + \phi_2 p^m, & \text{if } y = m, \\ 0, & \text{otherwise} \end{cases} \\ &= \left[\phi_0 + \phi_2(1-p)^m\right] I(y=0) + \phi_2 \binom{m}{y} p^y (1-p)^{m-y} I(0 < y < m) \\ &\quad + (\phi_1 + \phi_2 p^m) I(y = m), \end{aligned} \tag{1.1}$$

where $\phi_0 \in [0, 1)$ and $\phi_1 \in [0, 1)$ respectively denote the unknown proportions for incorporating extra zeros and extra right endpoints (or binomial denominators) than those allowed by the standard binomial distribution, and $\phi_2 \hat{=} 1 - \phi_0 - \phi_1 \in (0, 1]$. The $\text{EIB}(\phi_0, \phi_1; m, p)$ is a mixture of two degenerate distributions Degenerate(0), Degenerate(m) and a Binomial(m, p) distribution. In particular, when $\phi_0 = 0$, the EIB distribution is reduced to *right-endpoint inflated binomial* (REIB) distribution (denoted by $\text{REIB}(\phi_1; m, p)$); when $\phi_1 = 0$, the EIB distribution is reduced to the ZIB distribution (denoted by $\text{ZIB}(\phi_0; m, p)$); when $\phi_0 = \phi_1 = 0$, the EIB distribution becomes the standard binomial distribution.

The remainder of this paper is organized as follows. Section 2 provides six different but equivalent stochastic representations for the EIB random variable. Section 3 develops important distributional properties. In Section 4, we introduce the Fisher scoring algorithm and derive an *expectation–maximization* (EM) algorithm to find the *maximum likelihood estimates* (MLEs) of parameters in the model without any covariates. Bootstrap confidence intervals are also provided. In Section 5, generalized and fixed EIB regression models are proposed and the corresponding computational procedures are provided. In Section 6, we analyze a real data set. In Section 7, simulation studies are conducted to evaluate the performance of the proposed methods. A discussion is given in Section 8. All technical details are put in the supplemental document (see Appendix A).

2. Six different stochastic representations of the EIB random variable

In this section, we will establish six different but equivalent *stochastic representations* (SR) for the discrete random variable $Y \sim \text{EIB}(\phi_0, \phi_1; m, p)$.

2.1. Mixture of Degenerate(0), Degenerate(m) and Binomial(m, p)

Let $\mathbf{z} = (Z_0, Z_1, Z_2)^\top \sim \text{Multinomial}(1; \phi_0, \phi_1, \phi_2)$, $X \sim \text{Binomial}(m, p)$, and $\mathbf{z}$ and X be independent (denoted as $\mathbf{z} \perp\!\!\!\perp X$). We can show that the first SR of the random variable $Y \sim \text{EIB}(\phi_0, \phi_1; m, p)$ is given by

$$Y \overset{\text{d}}{=} Z_0 \cdot 0 + Z_1 \cdot m + Z_2 X = mZ_1 + Z_2 X = \begin{cases} 0, & \text{with probability } \phi_0, \\ m, & \text{with probability } \phi_1, \\ X, & \text{with probability } \phi_2, \end{cases} \tag{2.1}$$

Please cite this article in press as: Tian, G.-L., et al., Generalized endpoint-inflated binomial model. Computational Statistics and Data Analysis (2015), http://dx.doi.org/10.1016/j.csda.2015.03.009