Computational Statistics and Data Analysis 54 (2010) 1525–1534

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Unified computational methods for regression analysis of zero-inflated and bound-inflated data

Yan Yang[a,*], Douglas Simpson[b]

[a] *Department of Mathematics and Statistics, Arizona State University, Wexler Hall, Tempe, AZ 85287, USA*
[b] *Department of Statistics, University of Illinois, Illini Hall, 725 South Wright Street, Champaign, IL 61820, USA*

ARTICLE INFO

Article history:
Received 23 August 2008
Received in revised form 10 November 2009
Accepted 22 December 2009
Available online 4 January 2010

Keywords:
EM
Generalized estimating equation
Louis method
Mixture model
Quasi-Newton
Tobit model
Two-part model

ABSTRACT

Bounded data with excess observations at the boundary are common in many areas of application. Various individual cases of inflated mixture models have been studied in the literature for bound-inflated data, yet the computational methods have been developed separately for each type of model. In this article we use a common framework for computing these models, and expand the range of models for both discrete and semi-continuous data with point inflation at the lower boundary. The quasi-Newton and EM algorithms are adapted and compared for estimation of model parameters. The numerical Hessian and generalized Louis method are investigated as means for computing standard errors after optimization. Correlated data are included in this framework via generalized estimating equations. The estimation of parameters and effectiveness of standard errors are demonstrated through simulation and in the analysis of data from an ultrasound bioeffect study. The unified approach enables reliable computation for a wide class of inflated mixture models and comparison of competing models.

Published by Elsevier B.V.

1. Introduction

Bound-inflated data are prevalent in a wide variety of disciplines, such as health and safety studies, economics, finance and insurance risk analysis. Typically the responses are bounded below by zero with a significant mass of zero observations, resulting in data that are either discrete with too many zeros for a standard discrete distribution, or semi-continuous with positive continuous values combined with a substantial portion of zeros. In our collaborative research on ultrasound safety (O'Brien et al., 2006), groups of laboratory rabbits were exposed to focused ultrasound in both lungs to investigate the risk of ultrasound-induced hemorrhage. Due to the designed low to moderate acoustic pressure levels, about 80% of the observations were free of lesions while 20% exhibited lesions. The goal was to evaluate the effect of acoustic pressure and other factors based on the clustered, zero-inflated lesion size data to develop insight into the safe pressure levels for diagnostic clinical ultrasound.

A two-part model handles zeros and positive values, discrete or continuous, separately through two model components: a binary model for the occurrence of an event, and a zero-truncated Poisson or a log-normal model for the strictly positive size of the event conditional on its occurrence (Welsh et al., 1996; Zhou and Tu, 1999). For correlated counts with extra zeros, the zero-truncated Poisson and negative binomial models were extended by adding random effects to each model component (Yau and Lee, 2001; Min and Agresti, 2005). Dobbie and Welsh (2001) constructed generalized estimating equations (GEEs) with working correlation matrices for both components of the zero-truncated Poisson model. For semi-continuous longitudinal or clustered data, two-part random effects models were considered by Olsen and Schafer (2001)

[*] Corresponding author. Tel.: +1 480 965 6475; fax: +1 480 965 8119.
E-mail addresses: yy@math.asu.edu (Y. Yang), dgs@uiuc.edu (D. Simpson).

0167-9473/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.csda.2009.12.012

and Tooze et al. (2002). Albert and Shen (2005) proposed a two-part latent process model, which was recently adapted to incorporate random effects as well with Bayesian inference (Ghosh and Albert, 2009). For Bayesian two-part models with random effects, see also Zhang et al. (2006).

A zero-inflated (ZI) latent mixture model adds the point mass at zero to a discrete or censored distribution also capable of producing zeros. Two sub-models are involved: a binary model for the partially observed mixture-component indicators, and a Poisson regression as in the ZI Poisson model (Lambert, 1992) or a left-censored normal as in the ZI Tobit model (Cragg, 1971). A left-censored log-normal mixed with the point mass at a positive lower limit of detection was introduced by Moulton and Halsey (1995). In the presence of correlation, random effects were incorporated into either one or both sub-models (Hall, 2000; Berk and Lachenbruch, 2002; Yau et al., 2003; Lee et al., 2006). Alternatively, Moulton et al. (2002) implemented GEEs with the working independence correlation; Hall and Zhang (2004) developed a GEE approach for the class of ZI exponential family models.

Although much work has been done on fitting ZI data, most derivations have relied on special features of the individual models. We extend the existing latent mixture models through development of a unified framework, the left-inflated mixture (LIM) models for both discrete and semi-continuous data with point inflation at an arbitrary lower bound. This class not only includes current models but is broader by allowing various survival distributions (e.g., censored extreme value, logistic and t distributions) that add flexibility for modeling semi-continuous data. For correlated data, we construct GEEs with the working independence likelihood and estimate the covariance matrix of parameter estimates by the sandwich formula.

The quasi-Newton and EM algorithms are used for common estimation of model parameters. To find asymptotic standard errors associated with the EM, we investigate the generalized Louis method that extends the method of Louis (1982) to dependent data. For the quasi-Newton algorithm, a simulation study is conducted to assess the adequacy of estimating the outer Hessian matrix in the sandwich formula with the approximate Hessian at convergence. The performance and computational speed of the two methods are also compared empirically.

The rest of this article is organized as follows. Section 2 defines the left-inflated mixture models through a latent variable representation. Section 3 concerns maximum likelihood estimation for independent data and generalized estimating equation analysis for correlated responses. Section 4 discusses computational optimization of the estimating criteria by the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton and EM algorithms. Standard error estimation associated with each algorithm is discussed. Section 5 presents a simulation study that assesses and compares the two computational methods. The practical utility of our unified approach is illustrated with an ultrasound-induced lung hemorrhage study in laboratory animals in Section 6. Concluding comments are given in Section 7.

2. Left-inflated mixture models

Let $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_n^T)^T$ denote the multivariate response vector, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$ is the response vector for subject i, and Y_{ij} is the jth measure on subject i, $i = 1, \ldots, n, j = 1, \ldots, m_i$. The Y_{ij} are assumed to be bounded below (on the left) by L with a nonzero probability of observations equal to L. The lower boundary L is assumed to be known from the application under study or given by objective methods when a lower detection limit exists (Moulton and Halsey, 1995).

Let $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ be a realization of $\boldsymbol{Y}$. We consider left-inflated mixture models, in which the marginal distributions of the responses can be expressed as mixtures of distributions F_{ij} on $[L, \infty)$ and point masses concentrated at L. Here F_{ij} may be discrete or semi-continuous (as in the case of a left-censored distribution). The marginal densities have the form

$$\pi_{ij} f(y_{ij}) + \left(1 - \pi_{ij}\right) \delta_L(y_{ij}) = \begin{cases} 1 - \pi_{ij} + \pi_{ij} F_{ij}(L), & \text{if } y_{ij} = L \\ \pi_{ij} f(y_{ij}), & \text{if } y_{ij} > L \end{cases} \tag{1}$$

where $0 < \pi_{ij} \leq 1$ denotes the mixing weight, $F_{ij}(L) = F(Y_{ij} = L), f(\cdot)$ is the frequency or density function for F, and $\delta_L(u)$ equals one if $u = L$ and zero otherwise.

Such models have convenient latent variable representations. Define the mixture-component indicator vector $\boldsymbol{W} = (\boldsymbol{W}_1^T, \ldots, \boldsymbol{W}_n^T)^T$, where $\Pr(W_{ij} = 1) = \pi_{ij}$. Introduce a random vector $\boldsymbol{Z} = (\boldsymbol{Z}_1^T, \ldots, \boldsymbol{Z}_n^T)^T$ whose marginal distributions match F_{ij} on $[L, \infty)$, but whose distributions on $(-\infty, L)$ are chosen for computational convenience. Finally, assume that W_{ij} and Z_{ij} are pairwise independent. Then

$$Y_{ij} \overset{\mathrm{d}}{=} \left(1 - W_{ij}\right) L + W_{ij} \left\{L \cdot I(Z_{ij} \leq L) + Z_{ij} \cdot I(Z_{ij} > L)\right\} \tag{2}$$

for $i = 1, \ldots, n, j = 1, \ldots, m_i$, where "d" denotes in distribution and $I(A)$ is the indicator function for event A. If $L = 0$, then Eq. (2) yields a simplified representation for zero-inflated responses: $Y_{ij} \overset{\mathrm{d}}{=} W_{ij} Z_{ij} \cdot I(Z_{ij} > 0)$. For example, a ZI binomial random variable simplifies further to $W_{ij} Z_{ij}$ with $Z_{ij} \sim B(n_{ij}, \mu_{ij})$, while a ZI Tobit random variable has $Z_{ij} \sim N(\mu_{ij}, \sigma^2)$.

Consider regression models where both π_{ij} and μ_{ij} may depend on covariates through

$$\begin{aligned} h_1(\pi_{ij}) &= \boldsymbol{g}_{ij}^T \boldsymbol{\gamma} \\ h_2(\mu_{ij}) &= \boldsymbol{x}_{ij}^T \boldsymbol{\beta}, \end{aligned} \tag{3}$$