Computational Statistics and Data Analysis 128 (2018) 104–115

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Inference on zero inflated ordinal models with semiparametric link

Ujjwal Das [a], Kalyan Das [b,*]

[a] Operations Management, Quantitative Methods and Information Systems Area, Indian Institute of Management Udaipur, Udaipur 313001, Rajasthan, India
[b] Department of Mathematics, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India

ARTICLE INFO

Article history:
Received 23 June 2017
Received in revised form 28 March 2018
Accepted 28 June 2018
Available online 17 July 2018

Keywords:
Ordinal data
Zero inflation
Semiparametric regression
Knot selection
Sieve MLE

ABSTRACT

In socioeconomics or in Biological studies, observations on individuals are often observed longitudinally on a Likert-type scale with substantially large proportion of zeros. This leads to a special case of mixture structured data where extra-variation occurs. Obviously the standard ordinal data analysis fails to provide appropriate statistical inference. We propose a suitable zero inflated semiparametric ordinal model that takes into account the non linear link between the ordinal response and a covariate. A sieve maximum likelihood estimator(MLE) is proposed for the regression parameter of interest. We also propose a test for the zero proportion in this semiparametric model. A simulation study has been carried out to investigate the performance of the estimator as well as the test. We illustrate the methodology using data from a survey on Tuberculosis patients in and around Kolkata, India.

1. Introduction

Ordinal data are frequently encountered in a wide variety of disciplines ranging from economics to environmental studies. For example, insurance claim for a certain type of risk,transportation safety,severity of diseases , degree of pain, satisfaction with healthcare services are inherently ordinal in nature (Williamson et al., 1995; Gallefoss and Bakke, 2000). However often such data exhibit a high percentage of zeros (zero inflation) at the lower end of the ordinal scales, as compared to what is expected under an ordinal distribution. In drug abuse and treatment studies, there are considerable number of people who never use a particular drug (marijuana or cocaine for example) at all (none user), whereas others might not have used either drug at the time of survey (zero-consumption), but when the situation is conducive they may use the drug. The primary purpose of a zero inflated model is, therefore, to account for clumped zeros by incorporating sources of zero inflation: non-participation (structural zeros) or zero consumption (sampling zeros) (Mohri and Roark, 2005).

Zero inflation coupled with some multivariate structure makes it difficult to analyze the data and properly interpret the results. Methods that have been developed to address the zero-inflated data are limited to univariate-logit or univariate-probit model, Multinomial logit (McCullagh, 1980; Peterson and Harrell, 1990) or the ordered probit model (McKelvey and Zavoina, 1975; Greene and Hensher, 2010) are not appropriate for analyzing such clumped ordinal data. These conventional methods have limited capacity to explaining presence of zeros (Harris and Zhao, 2007). In contrast to the zero-inflated count data models where sufficient literature is available (see Lambert (1992), Gurmu and Trivedi (1996), and Mullahy (1997), Hall

* Corresponding author.
E-mail address: kalyan@math.iitb.ac.in (K. Das).

https://doi.org/10.1016/j.csda.2018.06.016
0167-9473/

(2000), Dagne (2004)), little attention has been paid to the problem of excess zeros in the ordered discrete choice models. To the best of our knowledge, the exception is the recent important paper by Harris and Zhao (2007), Gurmu and Dagne (2009) and Bagozzi et al. (2015). They develop a zero-inflated ordered probit model using the binary probit specification for the zero-inflation part of the model.

When such ordinal data of interest are of longitudinal by nature, it is natural to introduce a nonlinear function of time in the longitudinal model. The current investigation is motivated by a data on tuberculosis based on the survey conducted in hospitals in and around Kolkata, India. The ordinal outcomes are the bacterial load recorded longitudinally. Along with those observations, several concomitant characters like age, polymorphism, body mass index (BMI) etc. are recorded on each subject. As the age effects vary differentially in different age groups, it is reasonable to explore the possibility of nonlinear effect of age on outcomes through semiparametric regression analysis. Much research has been done on the estimation of the parametric component in a general framework, aiming to obtain asymptotically efficient estimators. Advances in statistical research on nonparametric function estimation, generalized linear models (GLMs) (McCullagh and Nelder, 1989), and generalized estimating equations (GEEs) (Liang and Zeger, 1986) have provided the impetus for developing semiparametric generalized partial ordinal models (GPOM, Das and Sarkar (2014)). Here "semiparametric" refers to the inclusion of a nonparametric covariate effect in an otherwise ordinal model.

In the present investigation our primary focus is to test whether the degenerate distribution at zero is necessary. If it is not, then no zero inflation needs to be modeled and the model simplifies merely to straightforward ordinal data analysis. Note that the presence of excess zero ensures extra variation in the data and simple ignorance may lead to bad estimation of parameters in the model. In this article we construct a semiparametric regression model in the presence of ordinal data with excess zero, and propose estimation of the parameters. We also develop a score test for determining the presence of zero inflation. We compare the behavior of the proposed test with likelihood ratio test through simulation study and apply our proposed approach to analyze a real data.

The organization of this paper is as follows. In Section 2, we discuss the zero inflated ordinal data model, and propose the semiparametric ordinal regression model. In the following subsections we discuss the parameter estimation for the proposed model. We then propose a score test for confirming the presence of excess zero in the data. In Section 3, we carry out a simulation study to investigate the suitability of the proposed model. In Section 4, we apply the proposed procedure to a tuberculosis data. Section 5 concludes with a brief discussion.

2. Methods

2.1. Model

Let y_i be an ordinal observation for ith subject, where $i = 1, 2, \ldots, n$ having levels $0, 1, 2, \ldots, J$ from a multinomial distribution : $y_i \sim$ *Multinomial*$(1, \pi_{i,0}, \pi_{i,1}, \ldots, \pi_{i,J})$ where $\pi_{i,j} = P(y_i = j)$ is the probability that ith individual falls in jth category so that $\pi_{i,0} + \pi_{i,1} + \cdots + \pi_{i,J} = 1$. Define the cumulative probability that ith individual falls in or below uth category as $P_{i,u} = P(y_i \leq u) = \sum_{k=0}^{u} \pi_{i,k}$ where $u \in \{0, 1, \ldots, J\}$. Here the zeros are assumed to appear from two ways. They either appear with a point mass ψ or they come from a multinomial distribution with probability $(1 - \psi)$. Sometimes the zeros with point mass are identified as structural zeros and those coming from the multinomial distribution are termed as sampling zeros. So, the entire data is assumed to come from a mixture distribution given by

$$P(y_i = y) = \begin{cases} \psi + (1-\psi)\pi_{0,i}, & \text{if } y = 0 \\ (1-\psi)(P_{i,j} - P_{i,j-1}), & \text{if } y = j \in \{1, 2, \ldots, J\} \text{ and } 0 \leq \psi \leq 1 \end{cases}, \tag{2.1}$$

For regression, let $\boldsymbol{x}_i$ denote the p-dimensional covariate vector for the ith subject. The parameter vector $(P_{i,0}, P_{i,1}, \ldots, P_{i,J})$ will be mapped with the predictors using a canonical link. Let $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{J-1})$ be the vector of intercepts and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ be the vector of regression coefficients. As a possible extension of the zero inflated ordinal model we consider a semiparametric link function. Then the model equation with partial linear link function along with an observable continuous predictor T may be defined as

$$\log \frac{P_{i,j}}{1 - P_{i,j}} = \gamma_j - x_i'\boldsymbol{\beta} + g(t_i) \tag{2.2}$$

with $j \in \{0, 1, 2, \ldots, (J-1)\}$. Here $g(.)$ is an unspecified smooth function for the effect of t. One may interpret γ_j as a constant representing the baseline value of the logit of cumulative probability for jth category and $\boldsymbol{\beta}$ is the effect of the covariates on them. $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, g, \psi)$ is the entire set of unknown parameters to be estimated. Let $Z = (y, X', T)$ be the data vector and $\boldsymbol{\theta}_0 = (\boldsymbol{\gamma}_0', \boldsymbol{\beta}_0', g_0, \psi_0)'$ as the true value of $\boldsymbol{\theta}$. In order to ensure the existence of the MLE we consider the following basic assumptions:

A.1: $\boldsymbol{\gamma} \in A_1$, $\boldsymbol{\beta} \in A_2$ where A_1 and A_2 are compact sets in $\mathcal{R}^J$ and $\mathcal{R}^p$. Further $\psi \in [0, 1]$, $T \in [0, 1]$ and X is such that $P(\|X\| \leq M) = 1$ for some constant M.

A.2: $S = \{g \in C^r[0, 1]; -\infty < m_0 \leq g(t) \leq M_0 < \infty\}\ \forall t \in [0, 1]$ and $r = 1,2$.