Computational Statistics and Data Analysis 96 (2016) 40–56

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Estimation and variable selection for proportional response data with partially linear single-index models☆

Weihua Zhao [a], Heng Lian [b,*], Riquan Zhang [c], Peng Lai [d]

[a] School of Science, Nantong University, Nantong, 226019, PR China

[b] School of Mathematics and Statistics, University of New South Wales, Sydney, 2052, Australia

[c] School of Finance and Statistics, East China Normal University, Shanghai, 200241, PR China

[d] School of Mathematics and Statistics, Nanjing University of Information and Technology, Nanjing, PR China

ARTICLE INFO

Article history:
Received 17 September 2014
Received in revised form 15 September 2015
Accepted 10 November 2015
Available online 29 November 2015

Keywords:
Estimating equation
Proportional data
Quasi-likelihood
Variable selection

ABSTRACT

Empirical researchers are often faced with the need to model proportional data in many fields such as econometrics, finance and biostatistics. In this paper, we study a robust and flexible modeling of proportional data using quasi-likelihood method with partially linear single-index structure. Bias-corrected estimating equations are developed to fit the model with the nonparametric function being approximated by polynomial splines. The theoretical properties of the estimators are established. In addition, we apply the regularization approach to simultaneously select significant variables and estimate unknown parameters, and the resulting penalized estimators are shown to have the oracle property. Extensive simulation studies and an empirical example are used to illustrate the usefulness of the newly proposed methods.

1. Introduction

In many empirical studies, researchers often focus on the variable of interest observed only on the standard unit interval, i.e. $0 \leq y \leq 1$, which is called proportional data or fractional data in the literature (Papke and Wooldridge, 1996; Ferrari and Cribari-Neto, 2004). Examples include the proportion of expenditure spent on food out of the entire family income, the proportion of exports in total sales, the proportion of debt in the financing mix of companies, the fraction of total weekly hours spent working, participation rates in voluntary pension plans, and the percentage of body fat, and so on. The main interests of these real examples are to find the relationships between proportional response variable and some important covariates.

The bounded nature of proportional data and, in some cases, the possibility of nontrivial probability mass accumulating at one or both boundaries raise some interesting issues. Since the usual estimation and inference tools are unsatisfactory for describing the proportional data, there exist some literatures on the proportional regression models in recent years. When there are no boundary values in the data set, i.e. the response variable y takes values on the open interval $(0, 1)$, Ferrari and Cribari-Neto (2004) proposed beta regression model to describe the data by the reparametrization of the beta response distribution, and thereafter many papers discussed related topics in the framework of beta regression, see Smithson

☆ The research was supported in part by National Social Science Fund of China (15BTJ027), National Natural Science Foundation of China (11171112) and Natural Science Fund of Nantong University (14B28).

* Corresponding author.
E-mail address: heng.lian@unsw.edu.au (H. Lian).

http://dx.doi.org/10.1016/j.csda.2015.11.004
0167-9473/

and Verkuilen (2006), Espinheira et al. (2008), Rocha and Simas (2011), Pereira et al. (2013), Zhao et al. (2014), and recent review paper of Ferrari (2013) and references therein. On the other hand, Song and Tan (2004) proposed to use simplex distribution to describe continuous proportional response data. The simplex distribution belongs to the family of dispersion models (Jøgensen, 1997), and there are several papers focusing on related estimation and inference problems based on the simplex distribution (Song et al., 2004; Song, 2009; Qiu et al., 2008; Zhang and Qiu, 2014). However, both beta regression and the simplex distribution based regression modeling need special distribution assumption for the response variable and they cannot directly deal with the case of proportional data with some boundary values. There are two approaches to deal with proportional data with some boundary values, one is to use zero- and/or one-inflated beta regression (Cook et al., 2008; Ospina and Ferrari, 2012), and the other is to take an ad hoc transformation to handle data at values of zero and one (Ramalho et al., 2011). However, there are some disadvantages for the aforementioned two methods, the former is very complicated and unstable in computation when the number of boundary values is not large, and the latter may result in inaccurate inferences and lack of reasonable interpretation for econometricians.

To circumvent these weaknesses of beta regression, in this paper, we developed a proportional response regression model with partially linear single-index structure based on quasi-likelihood method. Quasi-likelihood inference for proportional data was firstly proposed in the seminal paper of Papke and Wooldridge (1996), which only needs the specification of the first conditional moment of response variable and is a robust inference method without ad hoc transformation of boundary values. The quasi-likelihood method was further investigated in the econometric literature including its estimation methods, test issues, multivariate fractional data modeling with applications, see Ramalho et al. (2010), Ramalho et al. (2011), Murteira and Ramalho (in press) and Ramalho and Ramalho (2014). However, the above mentioned papers only focused on linear modeling, and there is little work on the nonparametric and semiparametric inference method for proportional data.

The partially linear single-index model (PLSIM) is a flexible and effective method that can avoid the problem of "curse of dimensionality" of multivariate modeling and capture the hidden nonlinear relationships between covariates and the response variable. PLSIM has been successfully investigated and applied to many kinds of data including cross-sectional data (Carroll et al., 1997; Yu and Ruppert, 2002; Liang et al., 2010; Lai and Wang, 2011), longitudinal data (Lai et al., 2013; Ma et al., 2014), and survival data (Lu and Cheng, 2007). However, there is no work on proportional data using PLSIM. In this paper, we study the estimation of PLSIM for proportional data based on quasi-likelihood. By using the polynomial splines to approximate the nonparametric function, we propose a profile estimation approach for the parametric component by constructing bias-corrected quasi-likelihood estimating equations and establish its theoretical properties. To further improve estimation, we develop methods to select the important variables based on the penalized bias-corrected quasi-likelihood estimating equations. Simulation results and real data analysis show that our newly proposed method has good performance.

Our proposed estimation procedure is established by the quasi-likelihood method, i.e., Bernoulli likelihood function. Therefore, our model can be seen as a special generalized PLSIM. In real data analysis, we do not suggest to use the simple method such as ordinary PLSIM for untransformed proportional data due to the bounded nature of proportional data, which lead to out of range estimates and predictions. On the other hand, we also do not use the Logistic transformation method or Logistic Normal distribution (Lesaffre et al., 2007) to investigate the PLSIM, i.e., use the pseudo-response $y^* = \log(y/(1-y))$ to obtain the estimates in the ordinary PLSIM. There are three main reasons: (1) When one uses the least squares method to fit the pseudo response variable y^*, if y^* is highly asymmetric, then the least squares based method may have low efficiency; (2) Interpretation of the regression coefficients becomes less straightforward; (3) There is no definition for proportional data with boundary values (0 and 1). For the last point, although Lesaffre et al. (2007) proposed to model the bounded outcome score on [0,1] by using the coarsened version of a latent score with Logistic Normal distribution, the proposed procedure seems harder to implement for complicated semiparametric models.

The rest of the paper is organized as follows. In Section 2, we introduce the quasi-likelihood function for proportional data with PLSIM and propose the bias-corrected quasi-likelihood procedure for estimation, where the nonparametric function is approximated by polynomial splines, and we further establish the large sample properties of the proposed estimators. To select important variables both in the nonparametric part and the linear part, we develop a variable selection procedure based on the double penalized bias-corrected quasi-likelihood estimating equations in Section 3, and their oracle properties are also investigated. Simulation studies in Section 4 and real data analysis in Section 5 are used to illustrate the performance of the proposed approach. Finally, some concluding remarks are given in Section 6. Technical proofs are deferred to Appendix.

2. Estimation methodology

2.1. Quasi-likelihood function

Suppose that $\{\boldsymbol{x}_i, \boldsymbol{z}_i, y_i\}_{i=1}^n$ are independent observations, where $0 \leq y_i \leq 1$ is associated with two covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iq})^T$, and n is the sample size. Usually, $\boldsymbol{x}_i$ are composed of continuous variables, while $\boldsymbol{z}_i$ are discrete. In this paper, the first conditional moment for response variable y_i is assumed as

$$\mathrm{E}(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i) = \mu_i, \quad i = 1, \ldots, n. \tag{1}$$