Computational Statistics and Data Analysis 99 (2016) 12–24

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Identification of proportionality structure with two-part models using penalization

Kuangnan Fang [a,b], Xiaoyan Wang [c], Ben-Chang Shia [d], Shuangge Ma [a,b,*]

[a] School of Economics, Xiamen University, China

[b] Department of Biostatistics, Yale University, United States

[c] College of Finance and Statistics, Hunan University, China

[d] School of Health Care Administration, Big Data Research Center & School of Management, Taipei Medical University, China

ARTICLE INFO

Article history:
Received 10 December 2014
Received in revised form 19 October 2015
Accepted 5 January 2016
Available online 13 January 2016

Keywords:
Zero-inflated data
Two-part modeling
Proportionality
Penalization

ABSTRACT

Data with a mixture distribution are commonly encountered. A special example is zero-inflated data, where a proportion of the responses takes zero values, and the rest are continuously distributed. Such data routinely arise in public health, biomedicine, and many other fields. Two-part modeling is a natural choice for zero-inflated data, where the first part of the model describes whether the responses are equal to zero, and the second part describes the continuously distributed responses. With two-part models, an interesting problem is to identify the proportionality structure of covariate effects. Such a structure can lead to more efficient estimates and also provide scientific insights into the underlying data-generating mechanisms. To identify the proportionality structure, we adopt a penalization method. Compared to the alternatives, notable advantages of this method include computational simplicity, solid statistical properties, and others. For inference, we adopt a bootstrap approach. The proposed method shows satisfactory performance in simulation and the analysis of two public health datasets.

1. Introduction

Data with a mixture distribution are commonly encountered. A special example is zero-inflated data. With such data, a proportion of the response values is equal to zero, and the rest of the responses have a continuous distribution. A motivating example is the household medical expenditure dataset analyzed in Section 4.1. This dataset was generated by the China Health and Nutrition Survey (CHNS) study, which was jointly conducted by the Carolina Population Center at the University of North Carolina at Chapel Hill, the National Institute of Nutrition and Food Safety, and the Chinese Center for Disease Control and Prevention. After standard data processing, it is observed that over 70% of the households had no medical expenditure during the study period, and the rest had continuously distributed expenditures (see Fig. 1 in the Appendix). In the literature, there are a large number of examples of zero-inflated data. See for example Cheung (2002), Agarwal et al. (2002), Deb et al. (2006), Bratti and Miranda (2011), Maruotti et al. (2015), and others.

Classic models assume that the zeros and the nonzeros in response come from the same data-generating process and may not be appropriate for data with excessive zeros. Multiple models have been developed to accommodate data withexcessive zeros. Notable examples include the hurdle models (Mullahy, 1986) which have been developed for count data with excessive zeros. Under the hurdle models, the two data-generating processes are not constrained to be the same. The basic idea

* Corresponding author at: Department of Biostatistics, Yale University, United States.
E-mail address: shuangge.ma@yale.edu (S. Ma).

http://dx.doi.org/10.1016/j.csda.2016.01.002
0167-9473/

is that a Bernoulli probability governs the binary outcome of whether a count variate has a zero or a positive value. If the value is positive, the hurdle is crossed, and the conditional distribution of the positives is governed by a truncated-at-zero count data model. Another family contains the zero-inflated models (Lambert, 1992), where the response variable is modeled as a mixture of a Bernoulli distribution (a point mass at zero) and a Poisson distribution (or another count distribution with support on non-negative integers). For data analyzed in this study and many others, the positive part has a continuous distribution. For such data, both the hurdle models and zero-inflated models are essentially the two-part models (Han and Kronmal, 2006; Liu et al., 2012). Under the two-part models, the first part describes whether theresponses take zero values (Manning et al., 1987; Olsen and Schafer, 2001). For those responses with nonzero values, the second part of the model describes their distribution. Notable advantages of the two-part models include intuitive interpretations, weak assumptions on the data generating mechanisms, and others. It is noted that in the recent literature, more complicated data structures have been considered. An example is longitudinal data, which have both within-subject correlation and between-subject heterogeneity and need to be accommodated using random effects (Min and Agresti, 2005; Greene, 2009; Alfó and Maruotti, 2010). For the aforementioned models, multiple estimation methods have been developed, including, for example, the quasi-likelihood method (McCulloch and Searle, 2001), penalized quasi-likelihood method (Yau and Lee, 2001), Bayesian method (Ghosh et al., 2006), and others. In most of the existing studies, the focus has been on modeling and estimation, while there is insufficient attention on the structure of covariate effects.

Consider covariate effects under the two-part modeling. The sets of covariates in the two model parts usually have a large overlap and, quite often, are identical. Even though the two parts have different formats, they in fact describe highly related underlying processes: the "growth" of a response from zero to nonzero (in a similar spirit as under the hurdle models), and from a small nonzero value to a large one. It is of interest to examine whether the two covariate effects in the two model parts are partially or completely proportional, that is, the proportionality structure of covariate effects. Research on theproportionality structure can be traced back to Cragg (1971). Lambert (1992) discusses the proportionality constraint for zero-inflated Poisson regression. Han and Kronmal (2006) proposes a hypothesis testing-based approach. Liu et al. (2011) develops a forward stepwise hypothesis testing-based approach and adopts bootstrap to compute the significance level of proportionality. Model selection-based approaches have also been developed. Liu and Chan (2011) develops a model selection criterion based on the marginal likelihood, which is similar to BIC. Liu et al. (2012) adopts a likelihood cross validation-based method.

Studying the proportionality structure is useful in multiple aspects. Statistically, if partial or full proportionality holds, then the model has a smaller number of unknown parameters than the unconstrained model, which can lead to improved efficiency (smaller variation) in estimation. This has been rigorously proved in Han and Kronmal (2006). Other studies, such as Liu et al. (2011), have demonstrated this using extensive numerical studies. Practically, proportionality may provide insights into the underlying data generating mechanisms. Consider for example the scenario with partial proportionality. Then the covariates can be separated into two classes: one has proportional effects in the two model parts and governs the two underlying data generating processes (from zero to nonzero, and from a small nonzero to a large one) in a similar manner, and the other behaves differently under the two processes. This can provide insights into the interconnections among covariates and their associations with the response.

Although the existing methods for determining the proportionality structure have achieved considerable successes, they also have limitations. Specifically, the hypothesis testing-based methods are computationally intensive. In addition, as sequential testing needs to be conducted, the control of type I error is nontrivial. The stepwise methods can be unstable (sensitive to even a small change in data). The BIC-based methods are computationally expensive when there are a moderate to large number of covariates. A common limitation shared by the existing methods is that computationally they do not "scale up" well. That is, their computational cost increases fast as the number of covariates increases. Most of the existing studies have focused on methodological development, and the statistical properties have not been well established.

In this study, we propose adopting penalization to determine the proportionality structure with two-part models. Penalization has been examined in a large number of studies. The goal of this study is not to develop a new penalization technique. Rather it is to apply penalization to a new statistical problem. The adopted method has an intuitive formulation as well as multiple technical advantages. It is computationally simple and can be realized using an effective algorithm. The computational cost increases relatively slowly with the number of covariates. Hence the proposed method can be applicable to data with a large number of covariates. Unlike in many of the existing studies, we take advantage of research on the asymptotics of penalization and rigorously establish the statistical properties, providing a solid ground to the proposed method. In addition, our numerical study suggests superior empirical performance of the penalization method.

2. Identification of proportionality structure using penalization

2.1. Two-part model and proportionality structure

Motivated by the data analyzed in Section 4, we consider the following distribution for the response variable Y

$$f(y) = (1-\phi)\mathbb{1}_{(y=0)} + [\phi \times N(y;\mu,\sigma^2)]\mathbb{1}_{(y>0)}, \quad y \geq 0,\ 0 \leq \phi \leq 1. \tag{1}$$

Under this model, there is a $1-\phi$ point mass at zero. For the positive part, motivated by the histograms in the Appendix, we adopt a normal distribution with mean μ and variance σ^2.