Computational Statistics and Data Analysis 103 (2016) 250–262

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A relative error-based approach for variable selection

Meiling Hao [a], Yunyuan Lin [b,*], Xingqiu Zhao [a]

[a] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China
[b] Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 7 October 2015
Received in revised form 6 February 2016
Accepted 23 May 2016
Available online 28 May 2016

Keywords:
Adaptive lasso
ADMM algorithm
Diverging number
Oracle property
Relative error

ABSTRACT

The accelerated failure time model or the multiplicative regression model is well-suited to analyze data with positive responses. For the multiplicative regression model, the authors investigate an adaptive variable selection method via a relative error-based criterion and Lasso-type penalty with desired theoretical properties and computational convenience. With fixed or diverging number of variables in regression model, the resultant estimator achieves the oracle property. An alternating direction method of multipliers algorithm is proposed for computing the regularization paths effectively. A data-driven procedure based on the Bayesian information criterion is used to choose the tuning parameter. The finite-sample performance of the proposed method is examined via simulation studies. An application is illustrated with an analysis of one period of stock returns in Hong Kong Stock Exchange.

1. Introduction

We consider the multiplicative regression model

$$Y_i = \exp(X_i^\top \boldsymbol{\beta})\varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where Y_i and X_i are pairs of response and p-vector of predictors, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is the p-vector of regression coefficient and ε is the positive unobservable random error. A multiplicative regression model or multiplicative error model is useful in analyzing data with positive response and heteroscedastic data, which are particularly common in economic, finance, reliability control, biomedical studies, epidemiological and social studies, etc. The estimation of model (1) is usually carried out by transforming the multiplicative models into linear models. However, a linear relationship in the transformed model is not linear in the original model. The analysis results based on linear models need to be transformed back to the original multiplicative measurement scale. A major aspect in regression analysis is variable selection or model selection. A variety of remarkable advancements have been developed for model selection; see for example, Chen and Donoho (1994), Tibshirani (1996), Fan and Li (2001), Shen and Ye (2002), Zou (2006), Wang and Leng (2007), Wang et al. (2007a,b), Wu et al. (2007), Huang et al. (2008a,b), Fan et al. (2009), Pötscher and Schneider (2009), Zhang and Lu (2007), Xu and Ying (2010), Huang et al. (2011), among many others.

The aforementioned estimation and model selection approaches are mostly based on criteria concerning the magnitude of absolute errors, for example, the least squares (LS) and least absolute deviation (LAD). However, in many practical applications, particularly in the analysis of heteroscedastic data, the LS and LAD methods are not adequate as they assign

* Corresponding author.
E-mail address: ylin@sta.cuhk.edu.hk (Y. Lin).

http://dx.doi.org/10.1016/j.csda.2016.05.013
0167-9473/

equal weights to the variables. For instance, in the analysis of a number of stocks, comparison of share prices of different stocks is generally meaningless, especially when there is possible share split or reverse split. In lifetime data analysis, longer life time requires more accuracy in terms of absolute error for prediction. In categorical data analysis, more accuracy for prediction in terms of absolute error may be required for a category with larger percentage of observations. There are a number of studies regarding relative errors in the literature; see Narula and Wellington (1977), Makridakis et al. (1984), Khoshgoftaar et al. (1992), Makridakis (1993), Park and Stefanski (1998), Chen et al. (2010), Gneiting (2011), Kolassa and Martin (2011), Zhang and Wang (2013), Tofallis (2014), Demongeot et al. (2015) and Chen et al. (2016), etc. In particular, Chen et al. (2010) proposed the least absolute relative error estimation for model (1) by taking two types of relative errors: one is the absolute error relative to the actual and the other is the absolute error relative to the predicted value of the target, into account in the parameter estimation simultaneously, which enjoys certain dimensionless/unitless and robust properties. Recently, in order to pursue a smooth and convex objective function incorporating relative errors, Chen et al. (2016) introduced a superior criterion called the least product relative error estimation (LPRE) to estimate $\boldsymbol{\beta}$.

As pointed out by Kolassa and Martin (2011) and Tofallis (2014), the most widely used measure for assessing prediction in business and organizations, the mean absolute percentage error or mean magnitude of relative error: the absolute error relative to the target, tends to select models whose prediction error is low. Similar consideration can be found in Demongeot et al. (2015) for a functional framework. A model selection approach, that would have advantages over existing methods in terms of interpretability and prediction accuracy, is much desired. To tackle the problem, in the present paper, we consider to borrow the ideas from Chen et al. (2016) and propose a statistical procedure based on product relative errors and Lasso-type penalties for variable selection and parameter estimation for multiplicative error models. First, the proposed procedure is based on two types of relative errors, which is symmetric in the actual and its predictor and therefore is a balanced and superior criterion compared with the commonly-used mean absolute percentage error. Second, the resultant estimator is dimensionless or scale-free, which retained the original measurement scale. Third, the smooth and convex nature of the LPRE allows numerical simplicity and ensures uniqueness of the solution. With certain proper choice of tuning parameters, the resulting estimator is proved to achieve the oracle property in both settings of fixed and diverging number of variables. The variance estimation can be carried out directly by a plug-in rule. An alternating direction method of multipliers (ADMM) algorithm is proposed for computing the regularization paths effectively. Furthermore, we adopt a BIC-type criterion to select the tuning parameter adaptively.

The rest of the paper is organized as follows. In Section 2, we introduce the proposed procedure along with a large sample theory including model selection consistency and the oracle property with fixed or diverging number of variables. Section 3 presents the ADMM algorithm to compute the resulting estimator. Section 4 reports some supportive simulation results and an application to a real dataset is given in Section 5. Some concluding remarks are given in Section 6 and all technical proofs are deferred to Appendix.

2. Methodology and main results

Let $(X_1^\top, Y_1), \ldots, (X_n^\top, Y_n)$ be n independent and identically distributed (i.i.d.) copies of $(X^\top, Y)$, where $X = (x_1, x_2, \ldots, x_p)^\top$ is the p-vector explanatory variable. Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$. Some notations are needed for ease of presentation. For two vectors $\boldsymbol{a} = (a_1, \ldots, a_d)^\top$ and $\boldsymbol{b} = (b_1, \ldots, b_d)^\top$, we define $\boldsymbol{a}\cdot\boldsymbol{b} = (a_1b_1, a_2b_2, \ldots, a_db_d)^\top$, $\boldsymbol{a}\cdot/\boldsymbol{b} = (a_1/b_1, \ldots, a_d/b_d)^\top$. Throughout the paper, the norm $\|\boldsymbol{a}\|_1 = \sum_{j=1}^d |a_j|$ and $\|\cdot\|$ is the Euclidean norm. For the multiplicative regression model, the least product relative error estimation proposed by Chen et al. (2016) is defined as the minimizer of

$$LPRE_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \left\{ \left| \frac{Y_i - \exp(X_i^\top\boldsymbol{\beta})}{Y_i} \right| \times \left| \frac{Y_i - \exp(X_i^\top\boldsymbol{\beta})}{\exp(X_i^\top\boldsymbol{\beta})} \right| \right\}.$$

Similar to the LARE in Chen et al. (2010), the LPRE accounts for two types of relative errors simultaneously, hence it is symmetric in the target and its predictor. The LPRE can also be regarded as the product of two weighted forms of absolute deviations. In the present paper, we propose a variable selection approach with the product relative errors loss and Lasso-type penalties. As pointed out by Fan and Li (2001) and Zou (2006), the Lasso does not achieve the oracle property in the sense that it cannot simultaneously set all unnecessary regression coefficients to zero correctly with probability tending to one as n increases while having the optimal rate of convergence. To obtain the oracle property, we consider to use the adaptive Lasso penalty in this paper; see Zou (2006), Wang et al. (2007a,b), Zhang and Lu (2007), among many others. A straightforward algebraic calculation of the LPRE criterion function yields

$$LPRE_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \left\{ Y_i \exp(-X_i^\top\boldsymbol{\beta}) + Y_i^{-1}\exp(X_i^\top\boldsymbol{\beta}) - 2 \right\}.$$

To be specific, we define the penalized LPRE estimator $\hat{\boldsymbol{\beta}}_n^*$ as the minimizer of

$$Z_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \left\{ Y_i \exp(-X_i^\top\boldsymbol{\beta}) + \exp(X_i^\top\boldsymbol{\beta})Y_i^{-1} \right\} + \lambda_n \|\boldsymbol{\beta}\cdot\boldsymbol{\omega}\|_1, \tag{2}$$