Journal of Statistical Planning and Inference xx (xxxx) xxx–xxx

ELSEVIER

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Frequentist model averaging estimation for the censored partial linear quantile regression model

Zhimeng Sun [a], Liuquan Sun [b,*], Xiaoling Lu [c], Ji Zhu [d], Yongzhuang Li [a]

[a] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China
[b] Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China
[c] School of Statistics, Renmin University of China, Beijing, China
[d] Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

A R T I C L E I N F O

Article history:
Received 23 March 2016
Received in revised form 24 April 2017
Accepted 25 April 2017
Available online xxxx

Keywords:
Model averaging
Model selection
Partial linear model
Quantile regression
Random censoring

A B S T R A C T

In this article, we propose a focused information criterion (FIC) and develop a frequentist model averaging estimation procedure for a partial linear regression model when the response is randomly right-censored. The proposed procedure is based on the quantile regression and can depict the comprehensive character of the distribution of the response by means of modeling different quantiles. The large sample properties of the proposed estimators are established, and their finite sample properties are examined through simulation studies. An application to a primary biliary cirrhosis data set is provided.

1. Introduction

The semiparametric partial linear regression model is a flexible generalization of the linear model and the nonparametric model. There has been a great amount of literature addressing the partial linear model from both theoretical and practical perspectives since it was first introduced by Engle et al. (1986). See, for example, Heckman (1986), Chen (1988) and Speckman (1988). Most of existing literature considers the partial linear model with fixed covariates. Obviously, model selection is also an important issue for a partial linear model, especially when researchers are able to collect richer data with the development of advanced techniques nowadays. To conduct model selection, several criteria including AIC (Akaike, 1973), BIC (Schwarz, 1978), Lasso (Tibshirani, 1997), SCAD (Fan and Li, 2001) and FIC (Claeskens and Hjort, 2003) can be employed. Nevertheless, as argued by many authors, these model selection procedures neglect the uncertainty in the selection process and may lose some useful information contained in potential models (e.g., Hjort and Claeskens, 2003; Yuan and Yang, 2005; Leung and Barron, 2006). An effective way to overcome the under-reporting problems of model selection procedures is by model averaging, rather than attaching to a single 'winning' model. Model averaging is a generalization of model selection and can provide a kind of insurance against model selection instability by weighting estimators across many potential models.

Much early work discussed model averaging techniques from a Bayesian perspective. Those methods are widely used recently due to the advance in computing techniques. However, model assumptions of these Bayesian model averaging

* Corresponding author.
E-mail addresses: sunzhimeng99@126.com (Z. Sun), slq@amt.ac.cn (L. Sun), xiaolinglu@ruc.edu.cn (X. Lu), jizhu@umich.edu (J. Zhu), liyongzhuang507@163.com (Y. Li).

http://dx.doi.org/10.1016/j.jspi.2017.04.001
0378-3758/

Please cite this article in press as: Sun, Z., et al., Frequentist model averaging estimation for the censored partial linear quantile regression model. J. Statist. Plann. Inference (2017), http://dx.doi.org/10.1016/j.jspi.2017.04.001

methods are complicated and hard to be visually explained. A useful overview of the literature is referred to Hoeting et al. (1999). Contributions from a frequentist perspective are fewer, but this strategy has received more and more attention in the recent years. For example, Buckland et al. (1997) proposed a smoothed BIC weight choice method for model averaging. Hjort and Claeskens (2003) suggested a smoothed FIC weight choice method. Hansen (2007) gave a weight choosing procedure through the Mallows' criterion. Zhang and Liang (2011) developed the smoothed FIC model averaging method for the generalized additive partial linear model. Zhang et al. (2014) considered the model averaging approach for the linear mixed-effects model. Xu et al. (2014) studied the focused information criterion and frequentist model averaging for the partial linear quantile regression model. These works provide insightful theoretical results and effective tools for practical applications. However, most of the existing literature considers uncensored data and there are few effective procedures for censored data that is very common in many applications.

Censored data often arise in economics, biomedicine, industry and many other fields. For example, duration data in econometrics are typical censored response data. In biomedicine, the survival time of a patient is usually censored. A rich body of work exists with respect to the regression analysis of censored response data (Koul et al., 1981; Lai et al., 1995; Bang and Tsiatis, 2002; Jin et al., 2003; Portnoy, 2003; Chen et al., 2005; Zeng and Lin, 2007; Wang and Wang, 2009; Shows et al., 2010; Du et al., 2013; Wang, Zhou and Li, 2013). For example, Bang and Tsiatis (2002) proposed a semiparametric procedure for estimating parameters in the median regression model using a weighted estimating equation. Chen et al. (2005) provided a rank estimation procedure to the partial linear model based on the Wilcoxon–Mann–Whitney estimating function. Wang and Wang (2009) suggested a locally weighted censored quantile regression approach by adopting the redistribution of mass idea. However, to the best of our knowledge, there is no existing work considering the frequentist model averaging for censored response data.

In this paper, we develop a FIC and model averaging procedure for a partial linear model with randomly right-censored response. Unlike the traditional criteria aiming at selecting a 'best' model for all parameters, the FIC is adaptive for different focus parameters. The covariates may affect the response very differently in different quantile levels. Thus, to depict the comprehensive character of the distribution of the response, we have to consider the influence of the covariates on the center of the response as well as their influence on other quantiles. In addition, outliers might have significant impact on either the least-square or likelihood-based methods. Also, we may confront heavy-tailed model errors. All of these motivate us to make model selection and run averaging in quantile regression to handle these problems.

The remainder of the paper is organized as follows. Section 2 describes the model framework and presents the estimation procedure under sub-models. Section 3 specifies the FIC and model averaging procedure as well as the confidence interval for the focus parameters. In Section 4, we develop a resampling method to estimate the asymptotic covariance matrix of the proposed estimators. Section 5 reports some numerical results from simulation studies for evaluating the proposed method. An application to the primary biliary cirrhosis (BPC) data set is provided in Section 6. Some concluding remarks are given in Section 7. Proofs of theorems are relegated to the Appendix.

2. Estimation procedure under sub-models

Our aim is to model the relationship between the response T with a continuous distribution function F and its affecting explanatory variables. When censoring is present, some observations of the response cannot be observed but are known to be no less than the censoring values while others are completely observed. Let C be the censoring time with a continuous survival function $G(t) = P(C > t)$. In what follows, for simplicity, we assume that C is independent of T and the explanatory variables. The methods developed can be generalized to allow for the dependence between C and the explanatory variables, and some discussion about this is given in Section 7.

Suppose that the true relationship between T and its explanatory variables can be described by the following partial linear model

$$T = X^\top \beta_0(\tau) + g_0(Z, \tau) + \varepsilon(\tau), \tag{1}$$

at a fixed quantile level $\tau \in (0, 1)$, where X is a $d \times 1$ vector of covariates linearly related to the response, Z is a covariate nonlinearly related to the response, ε is the model error with zero τth conditional quantile given X and Z, $\beta_0(\tau)$ is the d-dimensional coefficient vector at the τth quantile, $g_0(\cdot, \tau)$ is an unknown smooth function at the τth quantile. Although the "intercept" term does not appear in model (1), it is actually included in the functional component. For simplicity, we assume that Z is distributed on a compact interval $[0, 1]$. Also we suppress τ in $\beta_0(\tau)$ and $g_0(\cdot, \tau)$ for notational convenience. Let $\{T_i, C_i, X_i, Z_i, \varepsilon_i; i = 1, \ldots, n\}$ be independent replicates of $\{T, C, X, Z, \varepsilon\}$. Suppose that we observe $\{Y_i, \delta_i, X_i, Z_i; i = 1, \ldots, n\}$, where $Y_i = \min(T_i, C_i)$, $\delta_i = I(T_i \le C_i)$, and $I(\cdot)$ is the indicator function.

To estimate the functional component g_0, we can approximate g_0 by spline functions under mild smoothness assumptions. Let $\mathcal{T}_n$, with $0 = t_1 = \cdots = t_l < t_{l+1} < \cdots < t_{m_n+l} < t_{m_n+l+1} = \cdots = t_{m_n+2l} = 1$ be a sequence of knots that partition the closed interval $[0, 1]$ into $m_n + 1$ subintervals $I_i = [t_{l+i}, t_{l+i+1})$ for $i = 0, \ldots, m_n - 1$ and $I_{m_n} = [t_{m_n+l}, t_{m_n+l+1}]$, where m_n increases with the sample size n. Also let $\mathcal{S}_n(\mathcal{T}_n, l)$ be the space of polynomial splines on $[0, 1]$ of degree $l \ge 1$ with knots $\mathcal{T}_n$. Then, $\mathcal{S}_n(\mathcal{T}_n, l)$ consists of functions that are polynomials of degree l on each of the subintervals, and are $l - 1$ times continuously differentiable on $[0, 1]$ for $l \ge 2$. Under proper conditions on g_0 (e.g., Condition (C2) below), according to Corollary 4.10 of Schumaker (1981), we can approximate g_0 as

$$g_0(z) \approx B^\top(z)\alpha_0, \tag{2}$$