Journal of the Korean Statistical Society 39 (2010) 533–544

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Empirical likelihood for single-index varying-coefficient models with right-censored data

Zhensheng Huang

School of Mathematics, Hefei University of Technology, Hefei, Anhui, 230009, PR China

ARTICLE INFO

Article history:
Received 28 July 2009
Accepted 17 December 2009
Available online 19 January 2010

AMS 2000 Subject Classifications:
62G08
62G20

Keywords:
Backfitting algorithm
Confidence regions/intervals
Empirical likelihood
Right censoring
Single-index varying-coefficient models

ABSTRACT

This paper is concerned with an estimation procedure of a class of single-index varying-coefficient models with right-censored data. An adjusted empirical log-likelihood ratio for the index parameters, which are of primary interest, is proposed using a synthetic data approach. The adjusted empirical likelihood is shown to have a standard chi-squared limiting distribution. Furthermore, we increase the accuracy of the proposed confidence regions by using the constraint that the index is of norm 1. Simulation studies are carried out to highlight the performance of the proposed method compared with the traditional normal approximation method.

1. Introduction

In multivariate regression analysis, the so-called "curse of dimensionality" may occur without imposing some forms on the method. To attack this difficulty, many powerful modeling approaches have been incorporated to avoid the so-called "curse of dimensionality". An important example is the single-index model (e.g. Härdle, Hall, & Ichimura, 1993), which reduces the dimensionality from multivariate $\mathbf{X}$ to a univariate index $\beta^T\mathbf{X}$. More generally, Xia and Li (1999) suggested using a nonlinear index $\varphi(\beta, \mathbf{X})$, where φ is known up to a parametric vector β. In practice, usually a linear index is considered, i.e. $\varphi = \beta^T\mathbf{X}$ with $\|\beta\| = 1$. Therefore, we consider the following single-index varying-coefficient model (SIVCM)

$$Y = g^T(\beta^T\mathbf{X})\mathbf{Z} + \varepsilon, \tag{1}$$

where $Y \in R$ is a response variable, $\mathbf{X} = (X_1, \ldots, X_p)^T \in R^p$ and $\mathbf{Z} = (Z_0, Z_1, \ldots, Z_{q-1})^T \in R^q$ are covariates; $g(\cdot) = (g_0(\cdot), g_1(\cdot), \ldots, g_{q-1}(\cdot))^T$ is an unknown coefficient function vector. Without loss of generality, Z_0 can take a value identical to 1 so that $g_0(\cdot)$ is an intercept function. The error ε is independent of $(\mathbf{X}, \mathbf{Z})$ with $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. The unknown parametric vector $\beta = (\beta_1, \ldots, \beta_p)^T$ is in R^p and $\|\beta\| = 1$ for model identifiability.

The model (1) has considerable flexibility to cater for complex multivariate non-linear structure and includes many important models as special examples. For example, when $g_j(\cdot) = 0, j = 1, \ldots, q-1$, the model (1) reduces to the pure single-index models. If $g_j(\cdot), j = 1, \ldots, q-1$ are some constant parameters, (1) becomes the partially linear single-index models (e.g. Carroll, Fan, Gijbel, & Wand, 1997; Yu & Ruppert, 2002; Zhu & Xue, 2006). In addition, the model (1) also contains the standard varying-coefficient linear models (e.g. Cai, Fan, & Yao, 2000; Hastie & Tibshirani, 1993) with $g_0(\cdot) = 0$ and β is a known constant vector.

E-mail addresses: stahzs@126.com, my2008email@163.com.

1226-3192/$ – see front matter
doi:10.1016/j.jkss.2009.12.002

There are various approaches to the SIVCM (1). For example, under dependence, Xia and Li (1999) obtained the estimators of $g_j(\cdot), j = 0, \ldots, q-1$, by kernel smoothing and the estimator of β by the least squares method, and proved the asymptotic normality of the estimator of β and the consistency of the estimators for $g_j(\cdot), j = 0, \ldots, q-1$ with the optimal uniform convergence rate. Fan, Yao, and Cai (2003) discussed the local linear estimates of the model (1) with $p = q - 1$ and $X_j = Z_j, j = 1, \ldots, p$.

Obviously, the aforementioned approaches are used to model the relationship between the response and the predictor variables when the data are fully observable. However, when the response variable is censored the usual techniques are not directly applicable; i.e. for each $i = 1, \ldots, n$, the response variable Y_i is censored randomly on the right by some censoring variable C_i and cannot be observed completely. One only observe $\{(\mathbf{X}_i, \mathbf{Z}_i, \tilde{Y}_i, \delta_i), i = 1, \ldots, n\}$, where $\tilde{Y}_i = Y_i \wedge C_i$, $\delta_i = I[Y_i \leqslant C_i]$, and C_i is independent of $(\mathbf{X}_i, \mathbf{Z}_i, Y_i)$. To the best of our knowledge, this point has not been discussed in the literature. This motivates us to consider the single-index varying-coefficient models with right-censored data, which include two important statistical models: the censored partially linear single-index models proposed by Lu and Cheng (2007) and partially linear models with censored data discussed by Wang and Li (2002). And more discussion on censored regression models can be found in Lu and Cheng (2007).

In this paper, we make statistical inference for the parameters β in the single-index varying-coefficient models with right-censored data. Following the estimation procedure proposed by Lu and Cheng (2007), the least-squared estimates of β can be obtained and can be shown to be root-n consistent and asymptotically normal. Based on this, a normal-based confidence region for the parameters is constructed. However, under the censored data, the finite-sample performance of the normal-based confidence regions may not be appealing due to the complicated asymptotic covariance matrix. The empirical likelihood method introduced by Owen (1988) might be useful for the purpose of making semiparametric inference for the model (1). The empirical likelihood method has many advantages over normal approximation-based method. Firstly, the empirical likelihood does not involve a plug-in estimate for the asymptotic variance. Secondly, the empirical likelihood-based confidence region does not need to impose prior constraints on the region shape, and the region is range preserving and transformation respecting (see Hall & La Scala, 1990). Due to these nice properties, various applications of the empirical likelihood method can be found in Owen (1991), Chen and Hall (1993), Zhu and Xue (2006) and Huang and Zhang (2009). In addition, Owen (2001) proposed a fairly comprehensive reference.

To improve the accuracy of the normal-based confidences regions, we propose an alternative way for constructing confidence regions for the parameters β by using the empirical likelihood principle. We will show that the resulting empirical log-likelihood ratio is a mixture of central chi-squared distribution. To end this, an adjusted empirical log-likelihood ratio is proposed and is shown to have a central chi-squared limiting distribution. We also conduct simulation studies to compare the proposed method with the normal approximation method.

The rest of this paper is organized as follows. In Section 2 an adjusted empirical log-likelihood ratio is defined and some assumptions and main results are also given. Furthermore, the confidence regions for the parameters are constructed. Section 3 provides examples based on simulated data, and a comparison between the empirical likelihood method proposed and the normal approximation method is performed in term of coverage accuracy and areas/lengths of confidence regions/intervals. The proofs of the main results are collected in an appendix.

2. Methodology and main results

In this section we will derive an adjusted empirical likelihood method to make inference for the parameters β.

For the censored data, a difficulty for constructing confidence regions of β arises due to censoring. To end this, we use the synthetic data. In general, we can construct the following synthetic data

$$Y_{iG} = (1+\phi)L_{iG} - \phi K_{iG},$$

where $L_{iG} = \int_{-\infty}^{\infty}\{I[\tilde{Y}_i \geqslant s]/(1-G(s-)) - I[s < 0]\}ds$, $K_{iG} = \tilde{Y}_i\delta_i/(1-G(\tilde{Y}_i-))$, ϕ is a turning parameter which controls the weights put on the censored or uncensored observations, $G(\cdot)$ is the cumulative distribution function of the censoring time C_i. The choice of ϕ is very important in censoring regression. An appropriate choice of ϕ can reduce the variability of the transformed data. More discussion can be found in Fan and Gijbels (1994). In this paper, for notational simplicity we take $\phi = -1$, i.e. define a synthetic variable $Y_{iG} = \tilde{Y}_i\delta_i/(1-G(\tilde{Y}_i-))$, $i = 1, \ldots, n$, It can be verified that $E(Y_{iG}|\mathbf{X}_i, \mathbf{Z}_i) = E(Y_i|\mathbf{X}_i, \mathbf{Z}_i)$. Hence, under the model (1), we have

$$Y_{iG} = g^T(\beta^T\mathbf{X}_i)\mathbf{Z}_i + e_i, \quad i = 1, \ldots, n, \tag{2}$$

where $e_i = Y_{iG} - E(Y_{iG}|\mathbf{X}_i, \mathbf{Z}_i)$.

Following Zhu and Xue (2006), we will define an empirical log-likelihood ratio function as follows.

Because $\|\beta\| = 1$ means that the true value of β is the boundary point on the unit sphere, $g(\beta^T\mathbf{X}_i)$ does not have derivative at the point β. For this, we suggest the "delete-one-component" method proposed by Yu and Ruppert (2002). The details are as follows. We assume that the true parameter β has a positive component (otherwise, consider $-\beta$). Without loss of generality, we assume $\beta_r > 0$, where β_r is the rth component of β for $1 \leqslant r \leqslant p$. For $\beta = (\beta_1, \ldots, \beta_p)^T$, let