Computational Statistics and Data Analysis 88 (2015) 154–172

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Quantile regression methods with varying-coefficient models for censored data

Shangyu Xie [a,*], Alan T.K. Wan [b], Yong Zhou [c,d]

[a] RCAF and School of Banking and Finance, University of International Business and Economics, Beijing, China
[b] Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong
[c] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[d] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

ARTICLE INFO

Article history:
Received 11 July 2014
Received in revised form 12 February 2015
Accepted 14 February 2015
Available online 11 March 2015

Keywords:
Composite quantile regression
Covariate-dependent censoring
Inverse probability weighting
MM algorithm
Perturbation resampling

ABSTRACT

Considerable intellectual progress has been made to the development of various semi-parametric varying-coefficient models over the past ten to fifteen years. An important advantage of these models is that they avoid much of the curse of dimensionality problem as the nonparametric functions are restricted only to some variables. More recently, varying-coefficient methods have been applied to quantile regression modeling, but all previous studies assume that the data are fully observed. The main purpose of this paper is to develop a varying-coefficient approach to the estimation of regression quantiles under random data censoring. We use a weighted inverse probability approach to account for censoring, and propose a majorize–minimize type algorithm to optimize the non-smooth objective function. The asymptotic properties of the proposed estimator of the nonparametric functions are studied, and a resampling method is developed for obtaining the estimator of the sampling variance. An important aspect of our method is that it allows the censoring time to depend on the covariates. Additionally, we show that this varying-coefficient procedure can be further improved when implemented within a composite quantile regression framework. Composite quantile regression has recently gained considerable attention due to its ability to combine information across different quantile functions. We assess the finite sample properties of the proposed procedures in simulated studies. A real data application is also considered.

1. Introduction

Since the seminal work of Koenker and Bassett (1978), there has been an abundance of literature on various applications and theoretical extensions of quantile regression (QR). Regression quantiles have the important advantage over conditional mean regression of being able to directly estimate the effects of the covariates on quantiles other than the center of the distribution. It is also well-known that compared to the method of least-squares (LS), QR is more robust to outliers. QR has been extensively applied in economics, finance, biology, medicine, and many other disciplines. Recent empirical studies involving applications of QR can be found in Wheelock and Wilson (2008), Li et al. (2010), among others.

Although Koenker and Basset's (1978) conventional QR estimator is based on a linear parametric set-up, there has been a rapidly growing literature on the statistical theory and implementation of nonparametric and semiparametric QRs.

* Corresponding author.
E-mail address: xshyu@amss.ac.cn (S. Xie).

http://dx.doi.org/10.1016/j.csda.2015.02.011

For example, Koenker et al. (1994) discussed quantile smoothing splines; Yu and Jones (1998) considered nonparametric regression quantile estimation by kernel weighted linear fitting; De Gooijer and Zerom (2003) modeled the conditional quantile of the response as a nonlinear additive function of the covariates; and Wei and He (2006) developed a global QR approach to conditional growth charts. One important class of nonparametric models that has gained considerable attention in recent years is the varying-coefficient approach proposed by Cleveland et al. (1991) and Hastie and Tibshirani (1993). The appeal of the varying-coefficient model is that by allowing the coefficients to vary as smooth functions of other variables, the curse of dimensionality can be avoided. Due to this important advantage the varying-coefficient approach has experienced rapid development in theory and methodology. We refer to the articles by Cai (2007) for novel adaptations of the varying-coefficient approach to time series analysis; Fan and Li (2004) for longitudinal data analysis; Fan et al. (2006) and Cai et al. (2007, 2008) for survival analysis, and Wu et al. (2010) for functional linear regression. For more references, see Fan and Zhang (2008). To the best of our knowledge, Honda (2004), Kim (2007) and Cai and Xu (2008) are the only existing studies that consider the varying-coefficient approach for conditional quantiles; Honda (2004) and Cai and Xu (2008) used local polynomials to estimate conditional quantiles with varying coefficients, while Kim (2007) proposed an estimation methodology based on polynomial splines.

In recent years, we have also seen the emergence of a parallel literature on censored QR for which the usual set-up is one where the dependent variable of interest cannot be completely observed due to censoring. Censored QR was first studied by Powell (1986) for fixed censoring that assumes known censoring times for all observations. The examination of censored QR under the assumption of random censoring with unknown censoring points was taken up in a series of studies by Lindgren (1997), Yang (1999), Honoré et al. (2002), Gannoun et al. (2005) and Chen (2010). Portnoy (2003) studied censored QR under the self-consistency principle for the Kaplan–Meier estimator and developed a recursively re-weighted estimation procedure. Peng and Huang (2008) proposed a martingale-based estimating equations approach for censored QR models. This approach was subsequently extended by Qian and Peng (2010) to the analysis of a partially functional QR. Huang (2010) proposed a procedure for estimating censored QR based on estimating integral equations. The preponderance of this literature emphasizes nonparametric estimation of the conditional quantiles. However, some of the methods proposed in these studies rely on very strong distributional assumptions, or have major computational and/or theoretical drawbacks. For example, Lindgren's (1997) method requires an iterative minimization procedure for which no theoretical justification has been provided; this method is also computationally cumbersome when the dimension of the problem is high. The approach of Yang (1999) is restricted only to i.i.d. errors, and involves solving some highly complicated non-linear equations that can lead to multiple solutions. Portnoy (2003) also noted the algorithmic complications and computational issues associated with his proposed procedure. With few exceptions, most existing studies assume that censoring is independent of the covariates although it is not uncommon in practice to find correlations between the censoring time in the dependent variable and the covariates. For example, probability of loan default is typically thought to be associated with the borrower's credit worthiness reflected in the covariates. Dependence of censoring on covariates in QR is considered in the work of Portnoy (2003), Peng and Huang (2008) and Huang (2010), but none of these studies consider a varying-coefficient approach.

The major objective of the current paper is to develop a varying-coefficient approach to the estimation of regression quantiles under random data censoring when censoring times depend on covariates. We propose a weighted estimating function approach (Robins et al., 1994), whereby the contribution to the estimating function from an uncensored observation is weighted by the inverse of the probability of its being fully observed (Bang and Tsiatis, 2000). Inverse probability weighting is a widely used approach in censored data studies and has been adopted and refined in many subsequent studies (e.g., Zhao and Tian, 2001, Bang and Tsiatis, 2002 and Wang et al., 2012). One difficulty with this approach, however, is that the resultant estimating functions are non-smooth, rendering the Newton–Raphson algorithm inapplicable in solving the estimating equations. To reconcile this difficulty, we draw on the majorize–minimize (MM) algorithm (Hunter and Lange, 2000), and adapt it for the current analysis. A perturbation resampling method is developed for obtaining the estimator of the sampling variance. When censoring depends on the covariates, we can model this dependency either parametrically through, for example, the proportional hazards or the additive Aalen models, or nonparametrically by the Kaplan–Meier (Kaplan and Meier, 1958) estimator. Another purpose of this paper is to show that our procedure can be further improved when implemented within a composite quantile regression (CQR) framework. CQR was first introduced by Zou and Yuan (2008) for estimating coefficients in a linear regression. Extensions to local polynomial regression and varying-coefficient models were undertaken more recently by Kai et al. (2010, 2011). CQR has the appealing strength of combining information from different QRs, and hence the potential to improve estimation efficiency. All of the above-mentioned studies have shown that CQR can yield substantially more efficient estimators than LS-based procedures, but none of these studies have allowed for the possibility of censored data. This paper makes some progress towards obtaining results for CQR under censored data. Although this work may be thought of as an extension of Kai et al. (2011), the extension being considered is no means straightforward. Indeed, a very different set of analytical techniques is needed for obtaining results under random censoring using an inverse probability weighting approach. The efficiency gains of CQR over LS and QR in finite samples are examined by simulations.

The remainder of this paper is organized as follows. In Section 2, we present the model framework, and the local polynomial fitting and inverse probability weighting mechanisms, and describe the adaptation of the MM algorithm to the present analysis. In Section 3, we examine the asymptotic properties of the estimators, and provide a perturbation resampling method for variance estimation. Section 4 develops a CQR approach in the context of the censored varying-coefficient QR model. In Section 5, we report results of simulation studies that examine the performance of the proposed