Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

Estimation of general semi-parametric quantile regression

Q1 Yan Fan, Lixing Zhu *

Q3 *Renmin University of China and Hong Kong Baptist University, Hong Kong*

ARTICLE INFO

Article history:
Received 14 March 2012
Received in revised form
10 November 2012
Accepted 13 November 2012

Keywords:
Quantile regression
Single-index model
Outer-product gradient estimation (OPG)
qOPG
Eigenvector

ABSTRACT

Quantile regression introduced by Koenker and Bassett (1978) produces a comprehensive picture of a response variable on predictors. In this paper, we propose a general semi-parametric model of which part of predictors are presented with a single-index, to model the relationship of conditional quantiles of the response on predictors. Special cases are single-index models, partially linear single-index models and varying coefficient single-index models. We propose the qOPG, a quantile regression version of outer-product gradient estimation method (OPG, Xia et al., 2002) to estimate the single-index. Large-sample properties, simulation results and a real-data analysis are provided to examine the performance of the qOPG.

1. Introduction

Regression is used to quantify the relationship between a response variable Y and a p-variate covariate X, or specifically $F_{Y|X}$, the conditional distribution of Y given X. Unlike mean regression, which characterizes only the conditional mean of $F_{Y|X}$, quantile regression gives us a more complete picture of the relationship between the response variable Y and the covariate X. Quantile regression was first studied in linear quantile regression by Koenker and Bassett (1978) in their seminal work. Since then quantile regression has experienced deep and exciting developments in theory, methodology and applications. For example, a few nonparametric quantile regression models have been studied (Stone, 1977; Chaudhuri, 1991; Koenker et al., 1994; Yu and Jones, 1997, 1998). When some information on the form of the conditional quantile function of covariates is available, more efficient quantile regression models, such as semi-parametric model or nonparametric models with specific forms, are preferable (Gooijer and Zerom, 2003; Horowitz and Lee, 2005; Wu et al., 2010; Kong and Xia, 2011, etc.). For a more comprehensive review of quantile regression, see Koenker (2005) and references therein.

In this paper we propose a general semi-parametric quantile regression model. Suppose Y is a response variable and (X,Z) is a $(p+q)$-variate covariate with $X=(X_{(1)},\ldots,X_{(p)})^\top$ and $Z=(Z_{(1)},\ldots,Z_{(q)})^\top$. For $0<\tau<1$, we propose to model $Q_\tau(Y|X,Z)$, the τ quantile of Y given (X,Z), by a general semi-parametric model

$$Q_\tau(Y|X,Z)=G(X^\top\beta_0,Z), \tag{1}$$

where $G(\cdot,\cdot)$ denotes an unknown smooth function and β_0 is the unknown single-index coefficient. The dependence of G and β_0 is suppressed on τ as long as it does not cause an ambiguity.

* Corresponding author.
E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

0378-3758/$ - see front matter © 2012 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.jspi.2012.11.005

This semi-parametric model is rather general and flexible. It relates the covariate X to the conditional quantile of Y through a single-index $X^\top \beta_0$. This allows the dimension of X to be extremely large as the single-index can circumvent the so-called "curse of dimensionality" in high dimensional analysis. A nonparametric regression function is then involved to link the conditional quantile with the single index $X^\top \beta_0$ and other covariate Z. It includes many commonly used models as special cases, such as single index quantile regression models (Chaudhuri et al., 1997; Wu et al., 2010; Kong and Xia, 2011), partially linear quantile regression models (He and Shi, 1996; Lee, 2003), varying-coefficient quantile regression models (Honda, 2004; Kim, 2007; Cai and Xu, 2009) and partially linear varying coefficient quantile regression models (Wang et al., 2009; Kai et al., 2011), etc.

One important concern of model (1) is how to determine the covariates X and Z for given data. In general, there are two commonly used methods that can serve this purpose. If there are categorical or discrete explanatory variables in the data, a canonical partitioning of the explanatory variables is to take the continuous covariates as X, and the rest categorical or discrete covariates as Z, just as we did in our real-data analysis. In other cases, we can do the partition according to the meanings of the explanatory variables. For example, when analysis the circulatory and respiratory problems in Hong Kong, Xia and Hardle (2006) partitioned the explanatory variables into X and Z, according to whether they stand for weather conditions or air pollutants.

The focus of this paper is on the estimation of β_0. For model identification, we assume that $\|\beta_0\| = 1$ where $\|\cdot\|$ denotes the Euclid norm, and the first component of β_0 is positive. This problem was first considered by Koenker and Bassett (1978) in linear quantile regression, and then by Chaudhuri et al. (1997) in single-index quantile regression, in which the average derivative estimation (ADE) method was proposed. However the involvement of a high-dimensional kernel in ADE hinders its popularity. Another embarrassment of the ADE is that if the expectation of the derivative of the conditional quantile function with respect to β is zero, it fails to provide consistent estimate of β in theory. Wu et al. (2010) proposed a two-stage minimization procedure, which was shown to be more efficient than the ADE, to serve the same purpose in single-index quantile regression.

In this paper we propose a quantile regression based outer-product gradient (qOPG) estimation procedure to estimate β_0 in the general semi-parametric model. This idea is motivated by the outer-product gradient (OPG) method of Xia et al. (2002), which produces efficient estimates of the single-index coefficient in mean regression. The raw qOPG estimate is based on local linear estimates involving a high-dimensional kernel function. To alleviate the negative effect of the high dimension kernel, we take the raw qOPG estimate as an initial value and proposed a refined qOPG estimate using a lower dimensional kernel. We show that the qOPG estimates have root-n consistency, an optimal convergence rate as is achieved by the ADE estimate. Besides, one advantage of the qOPG estimates over the ADE estimates is that they works even if the derivative of the conditional quantile function with respect to X has mean zero, unless it is identically vanishing.

The rest of the paper is organized as follows. In Section 2, we present the proposed qOPG estimate, whose consistency is discussed in Section 3. Simulations are conducted in Section 4 to provide evidence for the effectiveness of the qOPG estimate. Regularity conditions and technical proofs are relegated to Appendix.

2. Methodology

2.1. The qOPG procedure

Suppose $G(u,z)$ is smooth enough. Let $G_1(u,z)$ and $G_2(u,z)$ denote the partial derivative of $G(u,z)$ with respect to u and z respectively. Then $\partial G(X^\top\beta_0,Z)/\partial X = G_1(X^\top\beta_0,Z)\beta_0$. The observation that all such partial derivatives are parallel to the single-index direction β_0 motivates us to estimate β_0 through estimating the partial derivative $\partial G(X^\top\beta_0,Z)/\partial X$ or $\mathsf{E}[\partial G(X^\top\beta_0,Z)/\partial X]$ (Chaudhuri et al., 1997).

A ready method for estimating $\partial G(X^\top\beta_0,Z)/\partial X$ is the local polynomial estimation procedure (Fan, 1992, 1993). Suppose that $\{(Y_i,X_i,Z_i),\ i=1,2,\ldots,n\}$ is a random sample from (Y,X,Z). For (X_i,Z_i) in the neighborhood of (X_j,Z_j), first-order Taylor approximation gives

$$G(X_i^\top\beta_0,Z_i) \approx a_j + \mathbf{b}_j^\top(X_i - X_j) + \mathbf{c}_j^\top(Z_i - Z_j),$$

where $a_j = G(X_j^\top\beta_0,Z_j)$, $\mathbf{b}_j = G_1(X_j^\top\beta_0,Z_j)\beta_0$ and $\mathbf{c}_j = G_2(X_j^\top\beta_0,Z_j)$. The local linear smoother defines the estimates of a_j, $\mathbf{b}_j$ and $\mathbf{c}_j$ by

$$(\hat{a}_j,\widehat{\mathbf{b}}_j,\widehat{\mathbf{c}}_j) = \arg\min_{a,\mathbf{b},\mathbf{c}} \sum_{i=1}^{n} \rho_\tau(Y_i - a - \mathbf{b}^\top X_{ij} - \mathbf{c}^\top Z_{ij}) w_{ij}, \tag{2}$$

where $\rho_\tau(u) = u(\tau - I(u<0))$ is the check function, $0<\tau<1$, and $w_{ij} \geq 0$ are some weights. Here and in what follows $X_{ij} = X_i - X_j$ and $Z_{ij} = Z_i - Z_j$. We shall discuss the choices of w_{ij} later.

Since all $\mathbf{b}_j$ are parallel to the single-index direction β_0, the direction of any linear combination of $\widehat{\mathbf{b}}_j$ or their average can be taken as a desirable estimate of β_0. This is the main idea of the ADE (Chaudhuri et al., 1997). However it fails when $\mathsf{E}G_1(X^\top\beta_0,Z)$ is equal or close to zero, because the average of $\widehat{\mathbf{b}}_j$ will also be close to zero and the information of β_0 in the average will be masked by random error. To overcome this dilemma and incorporate the information of all $\widehat{\mathbf{b}}_j$'s, we propose

Please cite this article as: Fan, Y., Zhu, L., Estimation of general semi-parametric quantile regression. Journal of Statistical Planning and Inference (2012), http://dx.doi.org/10.1016/j.jspi.2012.11.005