

# Statistical Modelling

<http://smj.sagepub.com>

---

## Comparison of kernel estimators of conditional distribution function and quantile regression under censoring

Ali Gannoun, Jérôme Saracco and Keming Yu  
*Statistical Modelling* 2007; 7; 329  
DOI: 10.1177/1471082X0700700404

The online version of this article can be found at:  
<http://smj.sagepub.com/cgi/content/abstract/7/4/329>

---

Published by:



<http://www.sagepublications.com>

**Additional services and information for *Statistical Modelling* can be found at:**

**Email Alerts:** <http://smj.sagepub.com/cgi/alerts>

**Subscriptions:** <http://smj.sagepub.com/subscriptions>

**Reprints:** <http://www.sagepub.com/journalsReprints.nav>

**Permissions:** <http://www.sagepub.in/about/permissions.asp>

**Citations** <http://smj.sagepub.com/cgi/content/refs/7/4/329>

# Comparison of kernel estimators of conditional distribution function and quantile regression under censoring

Ali Gannoun<sup>1</sup>, Jérôme Saracco<sup>2,3</sup> and Keming Yu<sup>4</sup>

<sup>1</sup> CNAM, Mathématiques CEDRIC, France

<sup>2</sup> Institut de Mathématiques de Bordeaux, Université Bordeaux 1, France

<sup>3</sup> GREThA, Université Montesquieu Bordeaux IV, France

<sup>4</sup> Department of Mathematical Sciences, Brunel University, UK

**Abstract:** We consider a regression model in which the variable of interest is censored. We present various nonparametric estimators of the conditional distribution function and of conditional quantiles. In a simulation study, we compare the performance of these estimators. Moreover, the local linear estimator of conditional quantile is applied on a dataset dealing with the effect of age on survival time of kidney transplant patients.

**Key words:** censored data; conditional quantile; kernel estimator; local linear estimator; survival analysis

Received July 2006; revised April 2007; accepted April 2007

## 1 Introduction

Let  $(X, T)$  be a random vector, where  $X$  is a random vector taking values in  $\mathfrak{R}^d$  and  $T$  is a real random variable. The coordinates of  $X$  may have various types of distributions, some of them may be discrete (for example binary), others may be absolutely continuous. Hence, we do not assume anything on the distribution of  $X$ . In this paper, we will consider conditional distribution function and regression quantile estimations in the presence of randomly right censored data. For  $\alpha \in (0, 1)$ , the conditional quantile  $q_\alpha(x)$  of  $T$  given  $X = x$ , also called regression quantile, is defined as the root of the following equation:

$$F_{T|X}(t|x) = \alpha,$$

where  $F_{T|X}(t|x) = P(T \leq t|X = x)$  is the conditional distribution function of  $T$  given  $X = x$ . Following Yu *et al.* (2003),  $q_\alpha(x)$  may be formulated as the solution to the

---

Address for correspondence: Jérôme Saracco, Université Bordeaux 1, Institut de Mathématiques de Bordeaux, UMR CNRS 5251, 351 cours de la libération, 33405 Talence Cedex, France. E-mail: Jerome.Saracco@math.u-bordeaux1.fr

following simple optimization problem:

$$q_\alpha(x) = \arg \min_{a \in \mathfrak{R}} E(\rho_\alpha(T - a) | X = x),$$

where  $\rho_\alpha(z) = z[\alpha - \mathbf{I}(z < 0)]$  and  $\mathbf{I}(\cdot)$  denotes the indicator function. Conditional or regression quantiles represent a substantially more general and informative method of regression analysis than conventional mean regression, since the former fully describe the conditional distribution of the response variable  $Y$  given  $X$ , without imposing any rigid distributional assumptions. Regression quantile estimation was introduced by Koenker and Bassett (1978) as a means of estimating conditional quantiles in linear regression models. It is widely used in a broad range of application settings. In paediatric medicine, reference growth curves for children's height and weight have a long history, and quantile regression methods may be used to estimate upper and lower reference curves as a function of age, sex and other covariates without imposing stringent parametric assumptions on the relationships among these curves; see Cole and Green (1992), and Gannoun *et al.* (2002). Koenker (2000) gives a survey of applications in economics in which conditional quantile is used to study determinants of wages, discrimination effects, and trends in income inequality, while Yu *et al.* (2003) provide a review from a statistics perspective. In ecology, theory often suggests how observable covariates affect limiting sustainable population size, and quantile regression can be used to directly estimate models for upper quantiles of conditional distribution rather than inferring such relationships from models based on conditional central tendency. In lifetime analysis, nonparametrically estimated conditional survival curves are useful for assessing the influence of risk factors, predicting survival probabilities and checking goodness-of-fit of various survival regression models; see Bowman and Wright (2000) for their interesting study on graphical exploration of covariate effect on survival time.

In many medical, industrial or economic studies, it is not possible to observe a sample of  $(X, T)$ . Hence, instead of an observation of  $T$ , one observes only the minimum of  $T$  and a censoring variable  $C$ . Then, we consider the problem of estimating the regression quantile function from such incomplete (or censored) data. To be more precise, we suppose that  $T$  is non-negative bounded random variable representing the survival or failure time of an individual or subject taking part in a medical or other experimental study. Assume that one observes  $n$  independent and identical distributed (i.i.d.) triples  $\{(X_i, Z_i, \Delta_i), i = 1, \dots, n\}$ , where  $Z_i = \min(T_i, C_i)$  and  $\Delta_i = \mathbf{I}(T_i \leq C_i)$ . The indicator random variable  $\Delta$  equals 1 if failure occurs before the censoring, and 0 otherwise. For identifiability, it is assumed that  $T$  is independent of  $C$  given  $X$ .

Nonparametric estimation, combining kernel and Kaplan-Meier (1958) methods are exhibited and studied by number of authors, and these successful application examples use a wide range of data sets in medicine, economics and social sciences; see among others, Beran (1981), Dabrowska (1992a), McKeague *et al.* (1995), Akritas (1994), Li and Doss (1995), Truong (2000), Leconte *et al.* (2002) or

Kohler *et al.* (2002) for random design regression, Van Keilegom and Veraverbeke (1998) or Bowman and Wright (2000) for fixed design function, and Cai (1998), Cai and Roussas (1998) or Leonenko and Sakhno (2001) under mixing framework (dependent data). Numerous other interesting works on censored quantile regression have been developed in literature, see for instance Chernozhukov and Hong (2002), Honore *et al.* (2002) or Portnoy (2003).

In the following, we review and improve some of the standard estimators of conditional distribution and conditional quantiles, and we point out some relations between them. Two typical kernel estimations of conditional distribution function are detailed in Section 2. In Section 3, we summarize several regression quantile estimators, including a local linear estimator. The smoothing selection rules for all these nonparametric estimators are discussed in Section 4. An extensive simulation study for the sample comparison of these estimators is carried out in Section 5. Finally, regression quantile estimation is used to analyze the data of kidney transplant patients in Section 6.

## 2 Estimation of the conditional distribution

As an alternative to the classical Cox (1972) regression model, Beran (1981) was the first who studied the estimation of conditional distribution without any parametric assumption. The main idea is the use of smoothing over the covariate space. Beran proposed two nonparametric estimators—one is based on nearest neighbor, and the other is based on kernel methods. He also proved the strong universal consistency in each case. In the following, we give an overview on Beran kernel estimator, and in the spirit of Cai (2003), we exhibit another one based on the kernel Weighted Least Squares (WLS) minimization.

### 2.1 Beran kernel estimator

In fact, Beran estimated the survival function  $S(t|x) = 1 - F_{T|X}(t|x)$ . There is no difficulty to deduce a kernel estimate of the conditional distribution, namely:

$$\widehat{F}_{T|X}(t|x) = 1 - \widehat{S}(t|x) = 1 - \prod_{Z_i \leq t, \Delta_i = 1}^n \left\{ 1 - \frac{\omega_{i,n}(x)}{\sum_{l=1}^n \omega_{l,n}(x) \mathbf{I}(Z_l \geq Z_i)} \right\},$$

with  $\omega_{i,n}(x) = K\left(\frac{x-X_i}{h_n}\right) / \sum_{l=1}^n K\left(\frac{x-X_l}{h_n}\right)$ . As regards  $\widehat{S}(t|X = x)$ , one can see Van Keilegom and Akritas (1999) or Li and Van Keilegom (2002) for instance. Dabrowska (1989) studied the asymptotic properties of  $\widehat{F}_{T|X}(t|x)$ . Further results, including bootstrap approximations, were obtained by Van Keilegom and Veraverbeka (1998).

## 2.2 Kernel WLS estimator

Even widely studied, Beran's estimator has the disadvantage that it does not behave well in the right tail when heavy censoring is present. To overcome this disadvantage, Gannoun *et al.* (2005) proposed an alternative estimator based on the local linear method developed by Cai (2003) to estimate the conditional expectation  $m(x) := E(T|X = x)$ . Let  $Z_{(1)} \leq \dots \leq Z_{(n)}$  be the order statistics of  $Z_1, \dots, Z_n$ , and let  $\Delta_{(i)}$  be the  $\Delta_i$  associated with the  $Z_{(i)}$ . To take censorship into account, let us introduce the following weight  $W_{[i]} = \frac{\Delta_{(i)}}{n-i+1} \prod_{l=1}^{i-1} \left( \frac{n-l}{n-l+1} \right)^{\Delta_{(i)}}$ . To estimate  $m(x)$ , Cai (2003) suggested the following least squares problem. Let  $X_{[i]}$  be the  $i$ th concomitant paired with  $Z_{(i)}$ , and let  $\hat{a}_0$  and  $\hat{a}_1$  minimize

$$\sum_{i=1}^n W_{[i]} \{Z_{(i)} - a_0 - a_1 (X_{[i]} - x)\}^2 \omega_{[i],n}(x).$$

Then  $m(x)$  is estimated by  $\hat{a}_0$ . In order to get an estimator of  $F_{T|X}(t|x)$ , one can notice that if we write  $Y = \mathbf{I}(T \leq t)$ , we have that  $E(Y|X = x) = F_{T|X}(t|x)$ . Then the estimation problem of the conditional distribution function may be viewed as a regression of  $Y$  on  $X$ . So, for fixed  $t$ , by minimization, with respect to  $b_0$  and  $b_1$ , of

$$\sum_{i=1}^n W_{[i]} \{\mathbf{I}(Z_{(i)} \leq t) - b_0 - b_1 (X_{[i]} - x)\}^2 \omega_{[i],n}(x), \quad (2.1)$$

we obtain  $\hat{b}_0 = \hat{F}_{T|X}(t|x)$ , the local linear estimator of  $F_{T|X}(t|x)$ , which can be expressed as

$$\hat{F}_{T|X}(t|x) = \frac{P_{n,2}(x)Q_{n,0}(x) - P_{n,1}(x)Q_{n,1}(x)}{P_{n,2}(x)P_{n,0}(x) - P_{n,1}^2(x)}, \quad (2.2)$$

where, for  $l \in \{0, 1, 2\}$ ,

$$P_{n,l}(x) = \sum_{i=1}^n \omega_{[i],n}(x)(X_{[i]} - x)^l W_{[i]} \quad (2.3)$$

$$\text{and } Q_{n,l}(x) = \sum_{i=1}^n \omega_{[i],n}(x)(X_{[i]} - x)^l W_{[i]} \mathbf{I}(Z_{(i)} \leq t). \quad (2.4)$$

It is easy to see that all the expressions (2.1), (2.2), (2.3) and (2.4) can be simplified by replacing  $\omega_{[i],n}(x)$  by  $K\left(\frac{x-X_{[i]}}{h_n}\right)$ .

Obviously, we can give another estimator of the conditional survival function by  $\widehat{S}(t|X = x) = 1 - \widehat{F}_{T|X}(t|x)$ . Because of their construction, asymptotic properties of  $\widehat{S}(t|X = x)$  and  $\widehat{S}(t|X = x)$  are similar to those of  $\widehat{F}_{T|X}(t|x)$  and  $\widehat{F}_{T|X}(t|x)$ .

### 3 Estimation of conditional quantiles

Our interest in this section is to estimate  $q_\alpha(x)$  under censoring. Survival analysis is concerned with the effect of a specific covariate on the survival time of an individual. A given covariate may have a different effect on low-, medium- and high-risk individuals. These effects can be understood by considering conditional quantile functions of survival time. Estimators can also be used to construct reference curves which are widely used in medicine and clinical studies.

#### 3.1 Nonparametric estimation of conditional quantile

Natural estimators of  $q_\alpha(x)$  can be deduced directly from the estimators  $\widehat{F}_{T|X}(t|x)$  or  $\widehat{F}_{T|X}(t|x)$  as the root of the equation  $\widehat{F}_{T|X}(t|x) = \alpha$  or  $\widehat{F}_{T|X}(t|x) = \alpha$ . We denote  $\widehat{q}_\alpha(x) = \widehat{F}^{-1}(\alpha|x)$  and  $\widehat{q}_\alpha(x) = \widehat{F}^{-1}(\alpha|x)$ .

Another way is to consider the local linear approach which can be essentially seen in Chaudhuri (1991) or Koenker *et al.* (1994). Our technique is similar to Cai (2003). By the use of elementary transformations, an estimator of conditional quantile is obtained by minimization of

$$\varphi_\alpha(x) = \sum_{i=1}^n \rho_\alpha(Z_i - c - d(X_i - x)) W_i K\left(\frac{X_i - x}{h_n}\right), \tag{3.1}$$

where  $W_i$  is the  $W_{[i]}$ -value associated with the natural order of  $Z_i$ . The local linear estimator of  $q_\alpha(x)$  is such that

$$\widehat{\widehat{q}}_\alpha(x) = \widehat{\widehat{c}}$$

where  $(\widehat{\widehat{c}}, \widehat{\widehat{d}})$  denotes the minimizer of (3.1). This method is direct and does not need the estimation of the conditional distribution. However, an obvious difficulty to get  $\widehat{\widehat{q}}_\alpha(x)$  is the absence of explicit representation. In addition, derivatives of  $\rho_\alpha$  do not exist everywhere. Following Lejeune and Sarda (1988), and Yu and Jones (1998), we propose an iteratively reweighted least square algorithm. Let us define new weights

$$\theta_\alpha(x, X_i, Z_i, c, d) = \begin{cases} \frac{\alpha}{Z_i - c - d(X_i - x)} & \text{if } Z_i - c - d(X_i - x) > 0 \\ \frac{\alpha - 1}{Z_i - c - d(X_i - x)} & \text{if } Z_i - c - d(X_i - x) < 0 \\ 0 & \text{if } Z_i - c - d(X_i - x) = 0 \end{cases}$$

and  $K_\alpha(x, X_i, Z_i, c, d) = \theta_\alpha(x, X_i, Z_i, c, d) W_i K\left(\frac{X_i - x}{h_n}\right)$ .

Then  $(\hat{c}, \hat{d}) = \arg \min_{c,d} \sum_{i=1}^n (Z_i - c - d(X_i - x))^2 K_\alpha(x, X_i, Z_i, c, d)$ . Now, make initial guesses  $(c_0, d_0)$  and then use the above formulation to iterate until convergence: if  $(c_l, d_l)$  are the values of  $(c, d)$  at the  $l$ th iteration, the next values  $(c_{l+1}, d_{l+1})$  will be given by

$$c_{l+1} = \frac{\sum_{i=1}^n K_\alpha(x, X_i, Z_i, c_l, d_l) (R_{n,2}(c_l, d_l) - (X_i - x)R_{n,1}(c_l, d_l)) Z_i^i}{R_{n,0}(c_l, d_l)R_{n,2}(c_l, d_l) - R_{n,1}^2(c_l, d_l)},$$

$$d_{l+1} = \frac{\sum_{i=1}^n K_\alpha(x, X_i, Z_i, c_l, d_l) (X_i - x)R_{n,0}(c_l, d_l) - R_{n,1}(c_l, d_l) Z_i}{R_{n,0}(c_l, d_l)R_{n,2}(c_l, d_l) - R_{n,1}^2(c_l, d_l)}.$$

Here,  $R_{n,l}(c, d) = \sum_{i=1}^n K_\alpha(x, X_i, Z_i, c, d)(X_i - x)^l$ ,  $l \in \{0, 1, 2\}$ . Practical choices of initial values will be discussed in a future section. Asymptotics of this local linear estimator have been studied in Gannoun *et al.* (2005).

An alternative algorithm is based on few efforts from the existing programs for a linear quantile model. For example, for each grid point  $x$ , the local polynomial quantile estimation can be implemented in the R package *quantreg* of Professor Koenker (<http://cran.r-project.org>) by setting covariates as 1 and  $(X_i - x)$ , and the weight as  $W_i K\left(\frac{X_i - x}{h_n}\right)$ .

### 3.2 Reference curves estimation

The conventional definition of reference range is a pair of numbers (the reference limits) that bind, for example, the central 90% of a set of values (called the reference values) obtained from a specified group of subjects (the reference subjects). The need for reference curves, rather than a simple reference range, arises when a covariate  $X$  is simultaneously recorded with the variable of interest  $Y$ . For  $\alpha > 0.5$ , reference curves are defined, when  $x$  varies, by  $I_\alpha(x) = [q_{1-\alpha}(x), q_\alpha(x)]$ . Clearly, estimation of reference curves is reduced to estimating conditional quantiles, namely we define  $\tilde{I}_\alpha(x) = [\tilde{q}_{1-\alpha}(x), \tilde{q}_\alpha(x)]$  as  $x$  varies, where  $\tilde{q}$  is  $\hat{q}$ ,  $\hat{\hat{q}}$  or  $\hat{\hat{\hat{q}}}$ .

### 4 Bandwidths selection

The main problem in the implication of nonparametric smoothing methods is the selection of the bandwidth in finite samples. Many data-driven bandwidth selection methods have been proposed for independent observations; see, for example, the very interesting survey article written by Jones *et al.* (1996). In the following, we discuss the selection of the parameter  $h_n$  via various methods, given the type of estimator. Notice that  $\tilde{R}^{-j}(U)$  is the leave-one-out estimator of  $R(U)$ .

#### 4.1 Beran estimator

Let us first give some additional notations:  $G(t|x) = P(C \leq t|x)$ ,  $H(t|x) = P(Z \leq t|x)$ ,  $H_1(t|x) = P(Z \leq t, \Delta = 1|x)$ ,  $F_X(x) = P(X \leq x)$  and  $f_X(x)$  its derivative.

The asymptotic mean squared error (AMSE) of this estimator is given by

$$AMSE \left( \hat{F}_{T|X}(t|x) \right) = h_n^4 b^2(t|x) + (nh_n)^{-1} s^2(t|x) \tag{4.1}$$

$$\begin{aligned} \text{where } b(t|x) &= \frac{1}{2} \int u^2 K(u) du (1 - F(t|x)) \left[ \int_{-\infty}^t \left\{ \frac{\ddot{H}(t|x) dH_1(t|x)}{(1-H(t|x))^2} + \frac{d\ddot{H}_1(t|x)}{1-H(t|x)} \right\} \right. \\ &\quad \left. + 2f'_X(x) f_X^{-1}(x) \int_{-\infty}^t \left\{ \frac{\dot{H}(t|x) dH_1(t|x)}{(1-H(t|x))^2} + \frac{dH_1(t|x)}{1-H(t|x)} \right\} \right], \\ s^2(t|x) &= f_X^{-1}(x) \int K^2(u) du (1 - F(t|x))^2 \int_{-\infty}^t \frac{d\dot{H}_1(t|x)}{(1-H(t|x))^2} \end{aligned}$$

with  $\dot{H}(t|x) = \frac{\partial}{\partial x} H(t|x)$ ,  $\ddot{H}(t|x) = \frac{\partial^2}{\partial x^2} H(t|x)$ , and similarly for  $\dot{H}_1$  and  $\ddot{H}_1$ , see Dabrowska (1992b) or Van Keilegom *et al.* (2001).

The bandwidth plays an essential role in the trade-off between reducing bias and variance. By minimization of (4.1) with respect to  $h_n$ , one can get an optimal choice for the bandwidth, namely

$$h_n = h_n(t, x) = \left( \frac{s^2(t|x)}{4b^2(t|x)} \right)^{1/5} n^{-1/5}.$$

In practice, we should estimate the functions  $s$  and  $b$  because they are unknown; see Dabrowska (1992b) for suitable estimators. However, we propose the following cross validation (CV) criterion which spares us the trouble of estimating the functions  $s$  and  $b$ :

$$CV(h) = \sum_{i=1}^n \sum_{l=1}^n \left\{ \mathbf{I}(T_i \leq T_l) - \hat{F}_{T|X}^{-i}(T_l|X_i) \right\}^2.$$

## 4.2 Kernel WLS estimator

Because this estimator is a result of square minimization criterion, to choose  $h_n$  we can apply the generalized cross validation (GCV). Our approach is slightly different from that used by Cai (2003) in order to estimate the conditional mean. The GCV selects the optimal global bandwidth  $h_n$  which minimizes

$$GCV(h) = \sum_{j=1}^n \sum_{i=1}^n \sum_{l=1}^m W_i \left\{ \mathbf{I}(Z_{(i)} \leq \tau_l) - \widehat{F}_{T/X}^{-j}(\tau_l | X_{[i]}) \right\}^2,$$

where  $\tau_1, \dots, \tau_m$  are  $m$ -values randomly chosen in  $[Z_{(1)}, Z_{(n)}]$ .

## 4.3 Conditional quantile estimators

Optimal bandwidth of the estimator derived from Beran (1981) is discussed in Li and Van Keilegom (2002). Concerning the WLS estimator  $\widehat{q}_\alpha(x)$ , we can use similar approach as in Li and Datta (2001), where the optimal bandwidth  $h_n$  is obtained by using bootstrap techniques. Therefore, we only focus on the local linear estimator  $\widehat{\widehat{q}}_\alpha(x)$ . Our methodology is very close to the one developed by Yu and Jones (1998), and Yu *et al.* (2003). It is based on the rule-of-thumb approach.

Let  $h_{mean}$  be the optimal choice of  $h_n$  obtained by CV criterion in order to estimate  $m(x)$ :  $h_{mean} = \arg \min_{h \in \mathfrak{N}} \sum_{j=1}^n \sum_{i=1}^n (T_i - \widehat{m}_h^{-j}(X_i))^2$ , where  $\widehat{m}_h(x) = \sum_{i=1}^n \omega_{i,n}(x) \frac{\Delta_i Z_i}{G_n(Z_i)}$  is a kernel estimator of  $m(x)$  obtained from all the data except  $Y_j$ , with  $G_n(t)$ , the Kaplan-Meier (1958) estimator of  $G(t) = P(C > t)$ , defined by

$$G_n(t) = \begin{cases} G_{1n}(t) = \prod_{i=1}^n \left( 1 - \frac{1 - \Delta_j^{(i)}}{n - i + 1} \right)^{\mathbf{I}(Z_{(i)} \leq t)} & \text{if } t < Z_{(n)}, \\ \lim_{t \rightarrow Z_{(n)}, t < Z_{(n)}} G_{1n}(t) & \text{if } t \geq Z_{(n)}. \end{cases}$$

One rule for selecting the bandwidth in the  $x$  direction simply modifies the bandwidth  $h_{mean}$  and can be implemented as  $h_n(\alpha) = h_{mean} \left[ \frac{\alpha(1-\alpha)}{\phi\{\Phi^{-1}(\alpha)\}} \right]^{-1/5}$ , where  $\phi$  and  $\Phi$  are standard normal density and distribution functions, respectively.

## 5 Simulations

In this section we will carry out a number of simulations to illustrate the numerical performance of the proposed estimators. We suppose that the theoretical distributions

of  $X$ ,  $T$  given  $X$ , and  $C$  given  $X$  are known, we proceed by the following way to get  $n$  observations  $\{(X_i, Z_i, \Delta_i), i = 1, n\}$  from  $(X, Z, \Delta)$ , in order to build the estimates. We first generate  $n$  observations  $X_1, \dots, X_n$  of  $X$  using classical (Monte Carlo) method. Then, we generate  $n$  observations  $U_1, \dots, U_n$  of  $U$ , the uniform variable in  $[0, 1]$ . The observations  $T_1, \dots, T_n$  are obtained by solving the equations  $F_{T|X}(t|X_i) = U_i$  for  $i = 1, n$ . Similarly, we obtain  $n$  realizations of  $C$  from conditional distributions  $F_{C|X}(\cdot|X_i)$ . Now, from each  $(X_i, T_i, C_i)$ , we can make a right censored observation  $(X_i, \min(T_i, C_i), \Delta_i) = (X_i, Z_i, \Delta_i)$ , and the estimators are based on these  $n$  right censored observations.

In the following, we assume that the covariate  $X$  is uniformly distributed on the interval  $[0, 1]$ . We studied on simulations the two examples (Exponential and Weibull distributions) which have been proposed by Van Keilegom *et al.* (2001), and Li and Van Keilegom (2002) in similar context. Because of their similarities, we only exhibit the results obtained in the Exponential case.

The conditional distribution of the response  $T$  given the covariate  $X = x$  is defined by  $(T|X = x) \sim \text{Exp}\left((a_0 + a_1x + a_2x^2)^{-1}\right)$ , while the censoring time  $C$  has the conditional distribution  $(C|X = x) \sim \text{Exp}\left((b_0 + b_1x + b_2x^2)^{-1}\right)$ . Because  $T$  and  $C$  are independent, it is obvious to see that  $(Z|X = x) \sim \text{Exp}\left((a_0 + a_1x + a_2x^2)^{-1} + (b_0 + b_1x + b_2x^2)^{-1}\right)$  and  $P(\Delta = 0|X = x) = \frac{a_0 + a_1x + a_2x^2}{a_0 + b_0 + (a_1 + b_1)x + (a_2 + b_2)x^2}$ . Here  $(a_0, a_1, a_2)$  and  $(b_0, b_1, b_2)$  are chosen in such a way that  $a_0 + a_1x + a_2x^2 > 0$  and  $b_0 + b_1x + b_2x^2 > 0$  for all  $x \in [0, 1]$ .

The rest splits into two subsections. The first contains the results of simulations corresponding to the conditional distribution estimation. We examine the performance of the estimators under different censoring. In the second subsection, we present simulations on conditional quantiles.

## 5.1 Study of conditional distribution estimators

We carried out the simulations for samples of size  $n = 200$ . The results are obtained by using 1000 simulations. We chose the biquadratic kernel function  $K(x) = (15/16)(1 - x^2)^2\mathbf{I}(|x| \leq 1)$ . Bandwidths are chosen using the previous section. It is obvious that  $F_{T|X}(t|x) = 1 - \exp\left(-\frac{t}{a_0 + a_1x + a_2x^2}\right)$ . Results are given for  $a_0 = 1$ ,  $a_1 = -5$ ,  $b_0 = 5$ ,  $b_1 = 2$  and  $b_2 = 8$ . The constant  $a_2$  was determined in such a way that the probability of censoring for  $x = 0.2, 0.5$  and  $0.75$  achieved the given value  $(0.25, 0.5, 0.75)$ . The different values of  $a_2$  as well as the corresponding distribution functions are summarized in Table 1.

**Table 1** Parameter  $a_2$  and corresponding conditional distributions

$P(\Delta = 0)$	$x$	$a_2$	$F_{T X}(t x)$	$F_{Z X}(t x)$
0.25	0.20	47.666	$1 - \exp(-0.524t)$	$1 - \exp(-0.69931t)$
	0.50	94.666	$1 - \exp(-t/27.166)$	$1 - \exp(-0.17011t)$
	0.75	11.407	$1 - \exp(-t/3.666)$	$1 - \exp(-0.36365t)$
0.50	0.20	143	$1 - \exp(-t/5.72)$	$1 - \exp(-0.34964t)$
	0.50	15.666	$1 - \exp(-t/8)$	$1 - \exp(-0.5382t)$
	0.75	38	$1 - \exp(-t/6.062)$	$1 - \exp(-0.1446t)$
0.75	0.20	429	$1 - \exp(-t/17.16)$	$1 - \exp(-0.2331t)$
	0.50	102	$1 - \exp(-t/24)$	$1 - \exp(-0.16667t)$
	0.75	63.555	$1 - \exp(-t/33)$	$1 - \exp(-0.12121t)$

In the following, we give  $F_{T|X}(t|x)$  in several points  $t_j$  which are the most representative with respect to theoretical distribution curve plots, and we evaluate the corresponding estimate  $\widehat{F}_{T|X}(t_j|x)$  and  $\widehat{\widehat{F}}_{T|X}(t_j|x)$  (see Tables 2 and 3).

Table 2 shows that under lower censoring (25%), both Beran estimator and kernel WLS estimator perform reasonably well for the whole covariate range. The former does slight better than latter for the middle to smaller  $x$  but not on the big value of  $x$ .

Table 3 is based on approximately 50% censoring and indicates that except middle values of  $t$ , kernel WLS estimator in general fits better for the range of covariates than Beran estimator.

**Table 2** Results for the 25% censoring case

(a) $P(\Delta = 0 0.2) = 0.25, F_{T X}(t 0.2) = 1 - \exp(-t/1.907)$										
$t_j$	0.2	0.4	1	1.5	2	2.5	3	4	5	8
$F_{T X}(t_j x)$	0.09	0.19	0.41	0.54	0.65	0.73	0.79	0.88	0.93	0.98
$\widetilde{F}_{T X}(t_j x)$	0.09	0.19	0.43	0.54	0.63	0.75	0.79	0.88	0.92	1.00
$\widehat{\widehat{F}}_{T X}(t_j x)$	0.09	0.21	0.42	0.54	0.64	0.72	0.78	0.88	0.95	1.00
(b) $P(\Delta = 0 0.5) = 0.25, F_{T X}(t 0.5) = 1 - \exp(-t/27.166)$										
$t_j$	1	2	5	7	10	20	30	40	60	80
$F_{T X}(t_j x)$	0.04	0.07	0.17	0.23	0.31	0.52	0.67	0.77	0.89	0.95
$\widetilde{F}_{T X}(t_j x)$	0.04	0.09	0.16	0.22	0.32	0.53	0.66	0.79	0.88	0.96
$\widehat{\widehat{F}}_{T X}(t_j x)$	0.15	0.06	0.35	0.28	0.32	0.58	0.68	0.79	0.88	0.99
(c) $P(\Delta = 0 0.75) = 0.25, F_{T X}(t 0.75) = 1 - \exp(-t/3.666)$										
$t_j$	0.5	2	3	4	6	7	8	10	14	20
$F_{T X}(t_j x)$	0.13	0.42	0.56	0.66	0.80	0.85	0.88	0.93	0.97	0.99
$\widetilde{F}_{T X}(t_j x)$	0.73	0.87	0.79	0.76	0.93	0.93	0.93	0.99	0.99	0.99
$\widehat{\widehat{F}}_{T X}(t_j x)$	0.14	0.56	0.61	0.69	0.85	0.87	0.87	0.95	0.99	1.00

**Table 3** Results for the 50% censoring case

(a) $P(\Delta = 0 0.2) = 0.5, F_{T X}(t 0.2) = 1 - \exp(-t/5.72)$										
$t_j$	0.5	1	2	4	6	8	12	14	16	20
$F_{T X}(t_j x)$	0.08	0.16	0.29	0.50	0.65	0.75	0.88	0.91	0.94	0.97
$\widehat{F}_{T X}(t_j x)$	0.07	0.18	0.32	0.53	0.68	0.82	0.94	0.94	0.94	0.99
$\widehat{\widehat{F}}_{T X}(t_j x)$	0.15	0.25	0.36	0.62	0.71	0.83	0.89	0.96	0.97	0.98
(b) $P(\Delta = 0 0.5) = 0.5, F_{T X}(t 0.5) = 1 - \exp(-t/8)$										
$t_j$	0.5	2	4	6	7	8	10	12	16	20
$F_{T X}(t_j x)$	0.06	0.22	0.39	0.53	0.63	0.63	0.71	0.78	0.86	0.92
$\widehat{F}_{T X}(t_j x)$	0.08	0.25	0.36	0.59	0.71	0.93	0.93	0.93	0.99	0.99
$\widehat{\widehat{F}}_{T X}(t_j x)$	0.08	0.21	0.42	0.61	0.93	0.93	0.93	0.93	0.93	0.93
(c) $P(\Delta = 0 0.75) = 0.5, F_{T X}(t 0.75) = 1 - \exp(-t/6.062)$										
$t_j$	0.5	2	4	6	8	9	10	12	16	20
$F_{T X}(t_j x)$	0.08	0.28	0.48	0.62	0.73	0.77	0.81	0.86	0.92	0.96
$\widehat{F}_{T X}(t_j x)$	0.17	0.45	0.56	0.69	0.72	0.75	0.79	0.94	0.99	0.99
$\widehat{\widehat{F}}_{T X}(t_j x)$	0.07	0.27	0.46	0.60	0.73	0.79	0.80	0.89	0.94	0.96

For heavy censoring case (75%), we find that kernel WLS estimator may lose advantage over Beran estimator for small value of  $x$  but not for big ones (the corresponding tables are not given here). However, both estimators do not find uncomfortable to fit the underlying distribution in this case.

To compare these two estimators, we use the mean squared error (MSE) defined by  $MSE = \frac{1}{m} \sum_{j=1}^m (\widetilde{F}_{T|X}(t_j|x) - F_{T|X}(t_j|x))^2$  where  $\widetilde{F}_{T|X}$  is  $\widehat{F}_{T|X}$  or  $\widehat{\widehat{F}}_{T|X}$ . We give the results for  $x = 0.2, 0.5$  and  $0.75$ . Using this criterion, none of the two estimators appears to be uniformly better; see Table 4.

**Table 4** MSE of  $\widehat{F}_{T|X}$  or  $\widehat{\widehat{F}}_{T|X}$

$x$	0.2	0.50	0.75
$\widehat{F}_{T X}(t x)$	0.003	0.011	0.065
$\widehat{\widehat{F}}_{T X}(t x)$	0.002	0.023	0.014

### 5.2 Study of conditional quantile estimators

From the conditional distribution function of  $T$  given  $X = x$ , it is easy to get the theoretical expression of conditional quantile by resolving the equation  $F_{T|X}(t|x) = \alpha$ . Then, for the Exponential distribution we get  $q_\alpha(x) = -(a_0 + a_1x + a_2x^2) \ln(1 - \alpha)$ .

In the simulation, we will take  $a_0 = 1$ ,  $a_1 = 5$ ,  $a_2 = 100$ ,  $b_0 = 20$ ,  $b_1 = -10$  and  $b_2 = 5$  as parameters for the Exponential distribution and its corresponding censoring variable.

We carry out simulations for samples of size  $n = 25, 50$  and  $200$ , with conditional quantiles evaluated on  $x = \frac{i}{10}$ ,  $i = 1, 9$  for  $\alpha \in \{0.1, 0.5, 0.9\}$ . Bandwidths are selected by simplified CV for the estimator derived from Beran's one ( $\hat{q}_\alpha(x)$ ), that is,  $h_n$  is the minimizer of  $\sum_{i=1}^9 \sum_{j=1}^n (\hat{q}_\alpha^{-j}(X_i, h_n) - q_\alpha(X_i))^2$ ,  $\hat{q}_\alpha^{-j}(X_i, h_n)$  denotes the estimator of  $q_\alpha(X_i)$  obtained without the observation  $Y_j$ . Note that in study with real data,  $q_\alpha(X_i)$  is replaced by  $\hat{q}_\alpha(X_i)$ . Table 5 summarizes the corresponding value of the bandwidth for each  $\alpha$ , each  $n$  and each distribution. Note that only results concerning estimators  $\hat{q}_\alpha(x)$  and  $\hat{\hat{q}}_\alpha(x)$  are given. Those of  $\hat{\hat{q}}_\alpha(x)$  are very similar to  $\hat{q}_\alpha(x)$ , but the calculation of optimal bandwidth takes more time. Some of our experiments on our computer with S-Plus environment show that for a sample size 200, computing Beran kernel estimation of conditional distribution is typically 40 seconds faster than computing the WLS conditional distribution.

**Table 5** Bandwidths obtained by cross validation criterion

$n$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
<i>Bandwidth for <math>\hat{q}_\alpha(x)</math> estimator</i>			
25	0.071	0.047	0.067
50	0.048	0.041	0.045
200	0.013	0.031	0.012
<i>Bandwidth for <math>\hat{\hat{q}}_\alpha(x)</math> estimator</i>			
25	0.056	0.049	0.056
50	0.019	0.016	0.019
200	0.010	0.009	0.011

To select the bandwidth  $h_n$  for the local linear estimator  $\hat{\hat{q}}_\alpha(x)$ , we can use rule-of-thumb method described in Section 4.3. We begin by the evaluation of  $h_{mean}$ . For the Exponential distribution, we use the following simplified CV criterion:  $h_{mean} = \arg \min_{h \in \mathbb{R}} \sum_{i=1}^9 \sum_{j=1}^n (m(X_i) - \hat{m}^{-j}(X_i, h))^2$ , with  $m(X_i) = a_0 + a_1 X_i + a_2 X_i^2$ . Finally, the rule-of-thumb method allows to get values for  $h_n$  by the following relationship:  $h_n(\alpha) = 1.24 h_{mean}$  for  $\alpha = 0.1$  or  $0.9$ , and  $h_n(\alpha) = 1.095 h_{mean}$  for  $\alpha = 0.5$ . Using the optimal selection of  $h_n$  given in Table 5, we evaluate  $\tilde{q}_\alpha(i/10)$  for  $i = 1, 9$  with  $\tilde{q}_\alpha = \hat{q}_\alpha$  or  $\hat{\hat{q}}_\alpha$ . Then, we consider the out-of-sample performances given by  $MSE = \frac{1}{9} \sum_{i=1}^9 (\tilde{q}_\alpha(i/10) - q_\alpha(i/10))^2$ . Simulations are done for  $n = 25, 50, 100, 500$ . Because of their similarities, only results for  $n = 25$  are presented in Table 6.

All the results seem to show that the local linear estimator appears to be uniformly better than  $\hat{q}_\alpha(x)$ .

**Table 6** Performance of the estimators for  $n = 25$

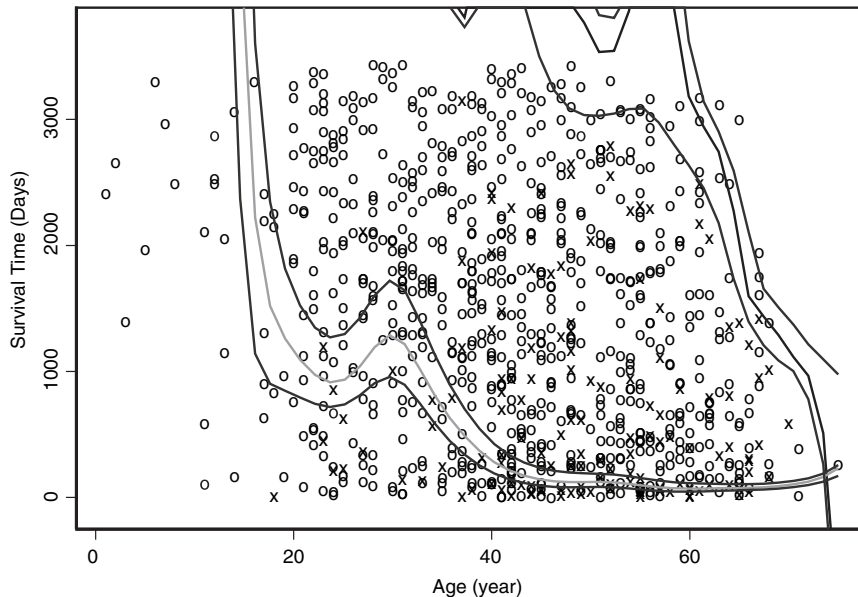
	$\alpha$	$\hat{q}_\alpha(x)$	$\hat{\hat{q}}_\alpha(x)$
<i>MSE</i>	0.1	1.086	0.945
	0.5	0.923	0.767
	0.9	1.092	0.874

## 6 Kidney transplant data analysis

Quantile regression under censoring provides an effective means of assessing the effect of a covariate on survival time. A given covariate may have different effects on low-, medium- and high-risk individuals. These effects can be understood by considering several quantile functions of survival time. The simulation study in Section 5 clearly indicates the superiority of local linear censored estimator  $\hat{\hat{q}}_\alpha(x)$ . To illustrate the practical performance of this estimator and the corresponding bandwidth selection rule, the data of kidney transplant patients will be used.

These data can be downloaded from the webpage of Professor John P Klein (<http://www.biostat.mcw.edu/homepgs/klein/kidtran.html>); they include 863 patients. We aim to investigate the effect of age on the long-term survival of kidney transplant patients. Figure 1 displays the data with the conditional median and 5% conditional quantile curves obtained by the local linear estimator and the proposed bandwidth selection. Because of over 80% observations being censored, any higher conditional quantiles than median ones lie beyond the range of observed values, so that it is impractical to build the standard reference charts mentioned in Section 3.2 for the survival time.

The conditional quantile curves displayed in Figure 1 are quite informative for the effect of the age of the patients. Clearly, the steep decline of conditional median curve for the age over 60 of the patients shows that for the majority of old patients the survival time is short. However, the 5% conditional quantile curve indicates that about 5% of the patient population has little change on survival after age over 40. These two curves also illustrate that, in spite of little variation in survival until age 60 for majority of patients, 5% of younger patients have decrease in survival until the age around 25 then increase to a local mode around the age of 30. Due to the presence of high censoring and high variability on survival time, these particular features or trends in the relationship between age and survival time may not be available without a complete description of underlying survival distribution, and



**Figure 1** The 50% (at the right-hand side) and 5% (at the left-hand side) quantile curves and their 90% confidence intervals for the kidney transplant patients, where o indicates the censoring points

quantile regression offers a such complete view of the effect of covariates on the location, scale and shape of the distribution of survival time. We are also be able to provide confidence intervals for the estimated conditional quantile curves in practice by using the asymptotic normality of Gannoun *et al.* (2005). Figure 1 displays the 90% confidence bands for the estimated two conditional quantile curves respectively.

## Acknowledgements

The authors are grateful to the editor and two anonymous referees for contributing to the improvement of this paper through useful remarks, suggestions and detailed comments.

## References

- Akritis MG (1994) Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics*, **22**, 1299–327.
- Berán R (1981) *Nonparametric regression with randomly censored survival data*, Technical report, University of California, Berkeley.
- Bowman AW and Wright E (2000) Graphical exploration of covariate effects on survival data through nonparametric quantile curves. *Biometrics*, **56**, 563–70.
- Cai Z (1998) Kernel density and hazard rate estimation for censored dependent data. *Journal of Multivariate Analysis*, **67**, 23–34.

- Cai Z (2003) Weighted local linear approach to censored nonparametric regression. In Akritas MG and Politis DM eds, *Recent advances and trends in nonparametric statistics*. Elsevier, 217–31.
- Cai Z and Roussas GG (1998) Kaplan-Meier estimator under association. *Journal of Multivariate Analysis*, **67**, 318–48.
- Chaudhuri P (1991) Nonparametric estimates of regression quantiles and their local Bahadur representation. *The Annals of Statistics*, **2**, 760–77.
- Chernozhukov V and Hong H (2002) Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association*, **97**, 872–82.
- Cole TJ and Green PJ (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305–319.
- Cox DR (1972) Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.
- Dabrowska DM (1989) Uniform consistency of kernel conditional Kaplan-Meier estimate. *The Annals of Statistics*, **17**, 1157–167.
- Dabrowska DM (1992a) Nonparametric quantile regression with censored data. *Sankhyā, Series A*, **54**, 252–59.
- Dabrowska DM (1992b) Variable bandwidth conditional Kaplan-Meier estimate. *Scandinavian Journal of Statistics*, **19**, 351–61.
- Gannoun A, Girard S, Guinot C and Saracco J (2002) Reference curves based on nonparametric quantile regression. *Statistics in Medicine*, **21**, 3119–135.
- Gannoun A, Saracco J, Yuan A and Bonney GE (2005) Nonparametric quantile regression with censored data. *Scandinavian Journal of Statistics*, **32**, 527–50.
- Honore B, Khan S and Powell JL (2002) Quantile regression under random censoring. *Journal of Econometrics*, **109**, 67–105.
- Jones MC, Marron JS and Sheather SJ (1996) Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*, **11**, 337–81.
- Kaplan E and Meier P (1958) Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association*, **53**, 457–81.
- Koenker R and Bassett GS (1978) Regression quantiles. *Econometrica*, **46**, 33–50.
- Koenker R, Ng P and Portnoy S (1994) Quantile smoothing splines. *Biometrika*, **81**, 673–80.
- Koenker R (2000) Galton, Edgeworth, Frish, and prospects for quantile regression in econometrics. *Journal of Econometrics*, **95**, 347–74.
- Kohler M, Máthé K and Pintér M (2002) Prediction from randomly right censored data. *Journal of Multivariate Analysis*, **80**, 73–100.
- Leconte E, Poiraud-Casanova S and Thomas-Agnan C (2002) Smooth conditional distribution function and quantiles under random censorship. *Lifetime Data Analysis*, **8**, 229–46.
- Lejeune MG and Sarda P (1988) Quantile regression: a nonparametric approach. *Computational Statistics and Data Analysis*, **6**, 229–39.
- Leonenko NN and Sakhno LS (2001) On the Kaplan-Meier estimator of long-range dependent sequences. *Statistical Inference for Stochastic Processes*, **4**, 17–40.
- Li G and Datta D (2001) A bootstrap approach to nonparametric regression for right censored data. *Annals of the Institute of Statistical Mathematics*, **53**, 708–29.
- Li G and Doss H (1995) An approach to nonparametric regression for life history data using local linear fitting. *The Annals of Statistics*, **23**, 787–823.
- Li G and Van Keilegom I (2002) Likelihood ratio confidence bands in nonparametric regression with censored data. *Scandinavian Journal of Statistics*, **29**, 547–62.
- McKeague IW, Nikabadze AM and Sun Y (1995) An omnibus test for independence of a survival time from a covariate. *The Annals of Statistics*, **23**, 450–75.
- Portnoy S (2003) Censored quantile regression. *Journal of the American Statistical Association*, **98**, 1001–12.

- Truong YK (2000) Asymptotics for hazard regression. Manuscript available on <http://www-bios.sph.unc.edu/~truong/man.html>
- Van Keilegom I and Veraverbeke N (1998) Bootstrapping quantiles in a fixed design regression model with censored data. *Journal of Statistical Planning and Inference*, **69**, 115–31.
- Van Keilegom I, Akritas MG and Veraverbeke N (2001) Estimation of the conditional distribution in regression with censored data: a comparative study. *Computational Statistics and Data Analysis*, **35**, 487–500.
- Van Keilegom I and Akritas MG (1999) Transfer of tail information in censored regression models. *The Annals of Statistics*, **27**, 1745–784.
- Yu K and Jones MC (1998) Local linear quantile regression. *Journal of the American Statistical Association*, **93**, 228–38.
- Yu K, Lu Z and Stander J (2003) Quantile regression: applications and current research areas. *The Statistician*, **52**, 331–50.