Journal of the Korean Statistical Society (2000), **29**: 1, pp 109–122

ADAPTIVE ESTIMATION OF EFFICIENT SCORE FUNCTION FOR CENSORED AND TRUNCATED REGRESSION MODELS †

Chul-Ki Kim [1]

ABSTRACT

An adaptive estimator of the efficient score function for censored and truncated regression models is developed by using B-splines to estimate the score function and a relatively simple cross validation method to determine the number of knots.

Keywords : Adaptation; B-splines; Cross validation; Censoring; Truncation.

1. INTRODUCTION

Consider the linear regression model

$$y_j = \beta^T x_j + \epsilon_j \quad (j = 1, 2, \cdots), \tag{1.1}$$

where the ϵ_j are i.i.d. random variables (representing unobservable disturbances) with a common distribution function F, β is a $d \times 1$ vector of unknown parameters and the x_j are either nonrandom or independent $d \times 1$ random vectors independent of $\{\epsilon_j\}$. Suppose that the responses y_j in (1.1) are not completely observable due to left truncation and right censoring by random variables t_j and c_j such that $-\infty \leq t_j < \infty$ and $-\infty < c_j \leq \infty$. It will be assumed that (t_j, c_j) are i.i.d. and are independent of (x_j, ϵ_j). Let $\tilde{y}_j = y_j \wedge c_j$ and $\delta_j = I_{\{y_j \leq c_j\}}$, where we use $\wedge$ and $\vee$ to denote minimum and maximum, respectively. In addition to right censorship of the responses y_j by c_j, we shall also assume left truncation in the sense that $(\tilde{y}_j, \delta_j, x_j)$ can be observed only when $\tilde{y}_j \geq t_j$. The data, therefore, consist of n observations $(\tilde{y}_i^o, t_i^o, \delta_i^o, x_i^o)$ with $\tilde{y}_i^o \geq t_i^o$, $i = 1, \cdots, n$. The special case $t_i \equiv -\infty$ corresponds to the "censored regression model" which is

† The research of Chul-Ki Kim is supported by the grant for promotion of scientific research in women's universities (97-N6-01-01-A-3).

[1] Department of Statistics, Ewha Womans University, Seoul, 120-750, KOREA.

of basic importance in statistical modelling and analysis of failure time data (cf. Kalbfleisch and Prentice (1980), Lawless (1982)). The special case $c_i \equiv \infty$ corresponds to the "truncated regression model" in econometrics (cf. Tobin (1958), Goldberger (1981), Amemiya (1985), Moon (1989)) and in astronomy (cf. Segal (1975), Nicoll and Segal (1980)), which assumes the presence of truncation variables τ_j so that (x_j, y_j) can be observed only when $y_j \leq \tau_j$ (or equivalently, when $-y_j \geq -\tau_j = t_j$). Left truncated responses that are also right censored arise in prospective studies of a disease and other biomedical studies (cf. Andersen et al.(1993), Keiding, Holst and Green (1989), Gross and Lai (1996)).

Lai and Ying (1991b,1992) studied efficient estimation of β from the data $(\tilde{y}_i^o, t_i^o, \delta_i^o, x_i^o)$ by developing asymptotic minimax bounds for the semiparametric estimation problem and constructing rank estimators that attain these bounds. Assuming that F has a continuously differentiable density function f so that the hazard function $\lambda = f/(1-F)$ is also continuously differentiable, their construction of these rank estimators consists of (i) dividing the sample into two disjoint subsets and evaluating a preliminary consistent estimate $\hat{b}_j$ of β from the jth subsample $(j = 1, 2)$, (ii) finding from the uncensored residuals in the jth subsample a smooth consistent estimate $\hat{\lambda}_j$ of the hazard function λ, (iii) smoothing $\hat{\lambda}_j'/\hat{\lambda}_j$ to obtain a smooth consistent estimate $\hat{\psi}_j$ of λ'/λ, and (iv) using $\hat{\psi}_1$ (respectively, $\hat{\psi}_2$) as the weight function for the linear rank statistic of the second (respectively, first) sample of residuals $\tilde{y}_i^o - b^T x_i^o$. The sum $S(b)$ of these two linear rank statistics is used to define the rank estimator as the minimizer of $\|S(b)\|$. There are, however, practical difficulties in carrying out this procedure.

First, rank estimators are difficult to compute when β is multidimensional. As noted by Lin and Geyer (1992), rank estimators of multidimensional β "require minimizing discrete objective functions with multiple local minima" and "conventional optimization algorithms cannot be used to solve such optimization problems." Computationally intensive search algorithms, such as the simulated annealing algorithm used by Lin and Geyer (1992), are needed to minimize $\|S(b)\|$. Another difficulty lies in estimation of λ'/λ to form the $\hat{\psi}_j$. Although there is an extensive literature on estimation of the hazard function λ and its derivative λ' for censored data, the problem of estimating λ'/λ from left truncated and right censored (l.t.r.c.) data is relatively unexplored. As will be shown in Section 2, simply plugging in $\hat{\lambda}_j'/\hat{\lambda}_j$ and smoothing the plugged-in estimate do not give good results unless the sample size is very large.

The present paper addresses these issues in constructing asymptotically efficient estimates of β from l.t.r.c. data. Instead of using rank estimators, we use

M-estimators which have much lower computational complexity (cf. Kim and Lai (1999)). These M-estimators are defined for l.t.r.c. data by the estimating equation

$$\sum_{i=1}^{n} x_i^o \{\delta_i^o \psi(\tilde{y}_i^o(b)) + (1-\delta_i^o) \int_{u>\tilde{y}_i^o(b)} \psi(u) d\hat{F}_b(u|\tilde{y}_i^o(b)) - \int_{u\geq t_i^o(b)} \psi(u) d\hat{F}_b(u|t_i^o(b)-)\} = 0, \tag{1.2}$$

where $\tilde{y}_i^o(b) = \tilde{y}_i^o - b^T x_i^o$, $t_i^o(b) = t_i^o - b^T x_i^o$, ψ is the score function associated with the M-estimator, and

$$\hat{F}_b(u|v) = 1 - \prod_{i:v<\tilde{y}_i^o(b)\leq u, \delta_i^o=1} \{1 - \Delta(b, \tilde{y}_i^o(b))/N(b, \tilde{y}_i^o(b))\}, \tag{1.3}$$

$$N(b,u) = \sum_{i=1}^{n} I(t_i^o(b) \leq u \leq \tilde{y}_i^o(b)), \Delta(b,u) = \sum_{i=1}^{n} I(\tilde{y}_i^o(b) = u, \delta_i^o = 1) \tag{1.4}$$

cf. Lai and Ying (1994). The notation $\hat{F}_b(u|v-)$ in (1.2) is used to denote (1.3) in which "$v < \tilde{y}_i^o(b)$" is replaced by "$v \leq \tilde{y}_i^o(b)$." The function $\hat{F}_b(u|-\infty)$ is the product-limit estimate of the common distribution function $F(u)$ of the ϵ_j in (1.1). Note that $\hat{F}_b(u|v)$ is the product limit estimate of

$$F(u|v) = P\{\epsilon_j \leq u | \epsilon_j > v\}. \tag{1.5}$$

Lai and Ying (1994) have shown that an asymptotically optimal choice of ψ in (1.2) is

$$\psi^* = (\lambda'/\lambda) - \lambda = f'/f, \tag{1.6}$$

for which the M-estimator of β is asymptotically normal with covariance matrix equal to that given by the information bound of the semiparametric estimation problem. Indeed this M-estimator of β has the same asymptotic properties as the asymptotically efficient rank estimator. Since the M-estimator has much lower computational complexity than the rank estimator, it will be better to be used for adaptive estimation in which the asymptotically efficient score function (1.6) is not assumed to be known *a priori* but has to be estimated from the data. Therefore, how well to estimate the efficient score function (1.6) from l.t.r.c. data will be the main factor for reducing computational complexity in the adaptive estimation. Throughout this paper, we focus attention on the problem of estimating the score function.

Section 2 discusses how (1.6) can be estimated. We use a spline approximation to ψ and a cross validation method to choose the number of knots. This is shown to perform much better than the plug-in method in Lai and Ying (1991b) based on estimating λ and λ'. Because of the simplicity of the cross validation method that involves only cross validating the two subsamples with each other, using this adaptive determination of the score function does not incur much increase in computational cost.

For complete data, Bickel (1982) showed how an adaptive estimate of β can be constructed so that it is asymptotically as efficient as the maximum likelihood estimate that requires specification of the density function f of the ϵ_j. The basic idea is to replace the unknown score function $(\log f)' = f'/f$ in the maximum likelihood estimate by $\hat{f}'_\eta/\hat{f}_\eta$, where $\hat{f}_\eta$ is a kernel estimate of f involving a bandwidth η that converges to 0 at a sufficiently slow rate as $n \to \infty$. Hsieh and Manski (1987) reported simulation studies showing that the behavior of an adaptive estimate can be changed dramatically in samples of moderate size by using different smoothing parameters η. They also proposed to choose the smoothing parameter that minimizes, over a preselected set of smoothing parameters, a bootstrap estimate of the mean squared error of $\hat{\beta}$. Faraway (1992) used B-splines to estimate $\log f$ so that the smoothing parameter is the number of knots (instead of the bandwidth in the kernel method) and estimated the mean squared error of $\hat{\beta}$ via an asymptotic formula instead of using the bootstrap. Jin (1992) used B-splines to estimate f'/f directly and proposed another cross validation method which we extend to l.t.r.c. data in Section 2, where alternative cross validation methods for l.t.r.c. data are also developed.

2. ESTIMATION OF THE EFFICIENT SCORE FUNCTION

In this section we assume known $\beta = 0$, so that the $y_j (= \epsilon_j)$ are i.i.d. with a common distribution function F that has a continuously differentiable density function f and hazard function λ, and consider estimation of the efficient score function (1.6) based on l.t.r.c. data. An obvious approach is to apply directly a method proposed by Uzunoğullari and Wang (1992) for estimating λ and λ' from l.t.r.c. data. The method estimates $\lambda^{(r)}(z)$ (the rth derivative for $r \geq 1$, with $\lambda^{(0)} = \lambda$) by $\hat{\lambda}^{(r)}(z) = \int K_{r,\eta}(z-u)d\hat{\Lambda}(u)$, where $\hat{\Lambda}$ is the estimated cumulative hazard function, $K_{r,\eta}(z) = \eta^{-(r+1)}K_r(z/\eta)$ and K_r is a kernel. The bandwidth $\eta (= \eta_{r,z})$ to estimate $\lambda^{(r)}(z)$ is chosen by a locally adaptive method that attempts to minimize the mean squared error of $\hat{\lambda}^{(r)}(z)$, replacing the unknown parameters

in the mean squared error by their estimates. Once $\hat{\lambda}$ and $\hat{\lambda}^{(1)}$ have been obtained by this procedure, one can estimate (1.6) by $\hat{\lambda}^{(1)}(z)/\hat{\lambda}(z) - \hat{\lambda}(z)$. However, this obvious estimate has difficulties when $\hat{\lambda}(z)$ is close to zero, as shown in Figure 1(a) where the estimate has a steep peak due to dividing by a small number. It is therefore natural to smooth out such peaks by applying some smoother to the estimate, but it is not clear how the smoothing parameter should be chosen. Alternatively, instead of estimating λ and λ' separately, one can apply numerical differentiation to $\log\hat{\lambda}(z)$ in estimating λ'/λ as is done in Figure 1(b), which also shows the need of smoothing the resultant estimate. The l.t.r.c. data in Figure 1 consist of 200 observations $(\tilde{y}_i^o, \delta_i^o, t_i^o)$ generated from the model of independent $\log(y_j) \sim N(0,1)$, $\log(t_j) \sim N(-1,1)$ and $\log(c_j) \sim N(1,1)$, with about 28% of the original sample being censored and truncated. The uncensored observations are represented by vertical bars along the horizontal axis. The kernels used in $\hat{\lambda}$

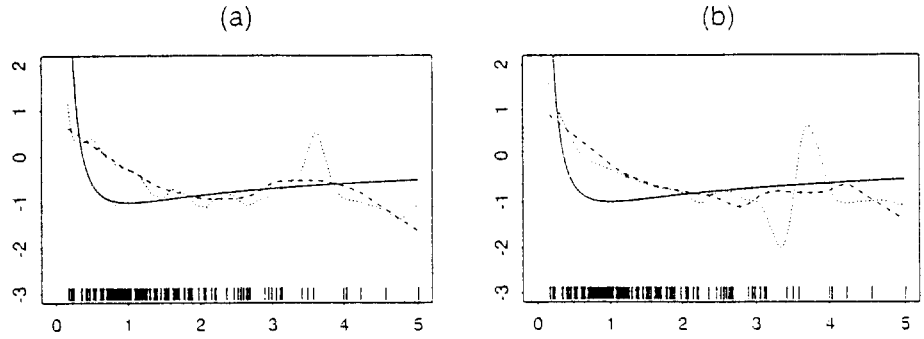

Figure1: (a) Left panel: the estimate $\hat{\lambda}^{(1)}/\hat{\lambda} - \hat{\lambda}$. (b) Right panel: the alternative estimate using numerical derivative of $\log(\hat{\lambda})$ to estimate λ'/λ. Both estimates are represented by dotted lines, while the solid line stands for the true score function. The broken curve is obtained by smoothing the dotted curve using Friedman's supersmoother.

and $\hat{\lambda}^{(1)}$ are

$$K_0(y) = \begin{cases} (15/16)(1-y^2)^2 & \text{if } |y| \le 1 \\ 0 & \text{otherwise} \end{cases},$$

$$K_1(y) = \begin{cases} (-15/4)y(1-y^2) & \text{if } |y| \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

Friedman's supersmoother (cf. Härdle (1990)) is used to smooth $\hat{\lambda}^{(1)}/\hat{\lambda} - \hat{\lambda}$ in Figure 1(a) and to smooth the numerical derivative of $\log\hat{\lambda}$ in Figure 1(b).

We next describe another method to estimate the efficient score function (1.6). First note that if f is the common density function of the $y_j (= \epsilon_j)$ then

$$f'/f = \lambda'/\lambda - \lambda. \tag{2.1}$$

We shall approximate $\psi^* = f'/f$ by B-splines, with knots located at certain quantiles of the product-limit estimate of F and with the number of knots chosen by cross validation for l.t.r.c. data.

2.1. Spline approximations to efficient score function

In an interval (a,b), take k knots $a < \xi_{k,1} < \cdots < \xi_{k,k} < b$, and define the linear B-spline basis $\{B_{k,i}\colon\ i = 0, \cdots, k+1\}$ as follows:

$$B_{k,i}(y) = \begin{cases} (\xi_{k,1} - y)/(\xi_{k,1} - a) & \text{if } a \le y \le \xi_{k,1} \text{ and } i = 0, \\ (y - \xi_{k,i-1})/(\xi_{k,i} - \xi_{k,i-1}) & \text{if } \xi_{k,i-1} \le y \le \xi_{k,i} \text{ and } 1 \le i \le k, \\ (\xi_{k,i+1} - y)/(\xi_{k,i+1} - \xi_{k,i}) & \text{if } \xi_{k,i} < y \le \xi_{k,i+1} \text{ and } 1 \le i \le k, \\ (y - \xi_{k,k})/(b - \xi_{k,k}) & \text{if } \xi_{k,k} < y \le b \text{ and } i = k+1, \\ 0 & \text{otherwise,} \end{cases}$$

where we set $\xi_{k,0} = a$ and $\xi_{k,k+1} = b$. We can define $B_{k,i}$ on the whole real line by setting $B_{k,i}(y) = 0$ for $y \notin [a,b]$. Let $D_{k,i}(y)$ be the derivative of $B_{k,i}$ at $y \notin \{\xi_{k,i-1}, \xi_{k,i}, \xi_{k,i+1}\}$ (or $y \notin \{a, \xi_{k,1}\}$ if $i = 0$, $y \notin \{\xi_{k,k}, b\}$ if $i = k+1$). Denote $B_k(y) = (B_{k,0}(y), \cdots, B_{k,k+1}(y))^T$, $D_k(y) = (D_{k,0}(y), \cdots, D_{k,k+1}(y))^T$, $A_k(y) = B_k(y)B_k^T(y)$, and

$$A_k(F) = \int_a^b A_k(y)dF(y),\ B_k(F) = \int_a^b B_k(y)dF(y),\ D_k(F) = \int_a^b D_k(y)dF(y) \tag{2.2}$$

Given the knots $a < \xi_{k,1} < \cdots < \xi_{k,k} < b$, the best linear spline approximation to ψ^* is defined as $a_k^T(F)B_k(x)$, where $a_k(F)$ minimizes $\int_a^b (a_k^T B_k(y) -$

$\psi^*(y)))^2 dF(y)$ over $a_k \in \mathbf{R}^{k+2}$. Since $\psi^* = f'/f$, integration by parts gives $\int_{-\infty}^{\infty} D_k(y) f(y) dy = \int_{-\infty}^{\infty} f dB_k = -\int_{-\infty}^{\infty} B_k \psi^* dF$, and therefore

$$\begin{aligned}
&\int_a^b (a_k^T B_k(y) - \psi^*(y))^2 dF(y) \\
&= a_k^T \left(\int_a^b A_k(y) dF(y) \right) a_k - 2a_k^T \int_a^b B_k(y) \psi^*(y) dF(y) + \int_a^b (\psi^*(y))^2 dF(y) \\
&= a_k^T A_k(F) a_k + 2a_k^T D_k(F) + \int_a^b (\psi^*)^2 dF, \qquad (2.3)
\end{aligned}$$

noting that B_k vanishes outside $[a, b]$. Since minimizing (2.3) is equivalent to minimizing $a_k^T A_k(F) a_k + 2a_k^T D_k(F)$, it follows that $a_k(F) = -A_k^{-1}(F) D_k(F)$.

In the case $k = 0$, we shall also denote the best linear approximation to ψ^* on $[a, b]$ by $a_0^T(F) B_0(y)$ to unify the notation, where we set $B_0(y) = (1, y - a)^T$.

2.2. Knot placement and extension of Jin's method for choosing the number of knots

Since F in (2.3) is unknown, we replace it by the product-limit estimate $\hat{F}$, which is defined in (1.3)–(1.4) with $v = -\infty$ and $(t_i^o(b), \tilde{y}_i^o(b))$ replaced by $(t_i^o, \tilde{y}_i^o)$. Take $0 < p < p^* < 1$ and set $a = \hat{F}^{-1}(p)$, $b = \hat{F}^{-1}(p^*)$. The knots $\xi_{k,i}$ $(i = 1, \cdots, k)$ are chosen to be the evenly spaced quantiles

$$\xi_{k,i} = \hat{F}^{-1}(p + (p^* - p) i/(k+1)). \qquad (2.4)$$

Ideally we would like to choose the number of knots to minimize $\int_a^b (a_k^T(\hat{F}) B_k(y) - \psi^*(y))^2 dF(y)$, or equivalently, to minimize $L(k, \hat{F}, F)$, where

$$L(k, G, F) = a_k^T(G) A_k(F) a_k(G) + 2a_k^T(G) D_k(F). \qquad (2.5)$$

Since F in (2.5) is unknown, one approach to implement the minimization of (2.5) with $G = \hat{F}$ is to extend Jin's method for complete data to the l.t.r.c. situation as follows:

1. Split the data into two subsamples $\{(\tilde{y}_1^o, \delta_1^o, t_1^o), \cdots, (\tilde{y}_{n_1}^o, \delta_{n_1}^o, t_{n_1}^o)\}$, $\{(\tilde{y}_{n_1+1}^o, \delta_{n_1+1}^o, t_{n_1+1}^o), \cdots, (\tilde{y}_{n_1+n_2}^o, \delta_{n_1+n_2}^o, t_{n_1+n_2}^o)\}$, where $n_1 = [n/2]$ and $n_2 = n - n_1$. Let $\hat{F}^{(1)}$ and $\hat{F}^{(2)}$ be the product-limit estimates based on these two subsamples separately.

2. Compute $L(k, \hat{F}^{(1)}, \hat{F}^{(2)}) = a_k^T(\hat{F}^{(1)}) A_k(\hat{F}^{(2)}) a_k(\hat{F}^{(1)}) + 2a_k^T(\hat{F}^{(1)}) D_k(\hat{F}^{(2)})$ for $k = 1, 2, \cdots$, and find the first local minimizer $\hat{k}_{cv}$ of $L(k, \hat{F}^{(1)}, \hat{F}^{(2)})$, i.e.

$$L(0, \hat{F}^{(1)}, \hat{F}^{(2)}) \geq \cdots \geq L(\hat{k}_{cv}, \hat{F}^{(1)}, \hat{F}^{(2)}) < L(\hat{k}_{cv} + 1, \hat{F}^{(1)}, \hat{F}^{(2)}).$$

3. Interchange $\hat{F}^{(1)}$ and $\hat{F}^{(2)}$ in Step 2, yielding the first local minimizer $\hat{k}'_{cv}$ of $L(k, \hat{F}^{(2)}, \hat{F}^{(1)})$.
4. Suppose $\hat{k}'_{cv} \leq \hat{k}_{cv}$ for definiteness. Compute $ST(k, \hat{F}) = k^{-1} \sum_{j=0}^{k-1} \int_a^b (a_j^T(\hat{F}) B_j(y) - a_k^T(\hat{F}) B_k(y))^2 \, d\hat{F}$ and find the first local minimizer $\hat{k}$ of $ST(k, \hat{F})$ over $k \in I(n)$, where $I(n) = \{k : \hat{k}'_{cv} \leq k \leq \hat{k}^2_{cv}\}$. Thus $ST(\hat{k}'_{cv}, \hat{F}) \geq \cdots \geq ST(\hat{k}, \hat{F}) < ST(\hat{k}+1, \hat{F})$. If there is no such $\hat{k}_n$ within $I(n)$, choose $\hat{k}_n = \hat{k}^2_{cv}$. This step is called "stationary correction" by Jin (1992), who explains its motivation as an attempt to ensure that $a_{k+1}^T(\hat{F}) B_{k+1}$ does not differ too much from $a_k^T(\hat{F}) B_k$ for the chosen k and thereby to reduce the variance of $\hat{k}$.

2.3. Cross validation for censored and truncated data

To begin with, note that an alternative way to combine $L(k, \hat{F}^{(1)}, \hat{F}^{(2)})$ and $L(k, \hat{F}^{(2)}, \hat{F}^{(1)})$ in Step 2 and 3 above is simply to add them so that $\hat{k}$ is defined as the minimizer of $L(k, \hat{F}^{(1)}, \hat{F}^{(2)}) + L(k, \hat{F}^{(2)}, \hat{F}^{(1)})$ over $0 \leq k \leq K_n$, some prescribed upper bound, instead of using Jin's stationary correction to combine the two subsample results. This is in fact tantamount to two-fold cross validation, as will be discussed below.

More generally, for m-fold cross validation, the dataset $\mathcal{S} = \{(\tilde{y}_1^o, \delta_1^o, t_1^o), \cdots, (\tilde{y}_n^o, \delta_n^o, t_n^o)\}$ is randomly divided into m disjoint subsets $\mathcal{S}_1, \cdots, \mathcal{S}_m$ with size $[n/m]$ for the first $m-1$ subsets and $n-(m-1)[n/m]$ for $\mathcal{S}_m$. Let $\hat{F}^{(\nu)}$ be the product-limit estimate of F based on $\mathcal{S}_\nu$ and let G_ν denote the product-limit estimate of F based on $\mathcal{S} - \mathcal{S}_\nu$. We use $\mathcal{S} - \mathcal{S}_\nu$ as the "training sample", from which estimates of the coefficients of the linear spline approximation are computed, and use $\mathcal{S}_\nu$ as the "test sample", leading to the measure $L(k, G_\nu, \hat{F}^{(\nu)})$ of the mean squared error (of using the training sample estimates to predict the efficient scores of the test sample values) minus $\int_a^b (\psi^*)^2 dF$, in view of (2.3). The m-fold cross validation approach chooses k to be the minimizer of $\sum_{\nu=1}^m L(k, G_\nu, \hat{F}^{(\nu)})$ over $k \leq K_n$. This way of defining m-fold cross validation requires n/m to be large enough so that $\hat{F}^{(\nu)}$ estimates F reasonably well. In the case of complete data, such requirement is actually not needed and one can in fact carry out full ("leave one out") cross validation with $m = n$, since $h(y_i)$ is an unbiased estimate of $\int_{-\infty}^{\infty} h dF$. Suppose h vanishes outside an interval (a, b). When y_i is not completely observable due to censoring and truncation, we can replace the

unobservable $h(y_i)$ by

$$h_F(\tilde{y}_i^o, \delta_i^o, t_i^o) = \delta_i^o h(\tilde{y}_i^o) + (1-\delta_i^o)\int_{\tilde{y}_i^o < y \le b} h(y)dF(y|\tilde{y}_i^o) + \int_{a \le y < t_i^o} h(y)dF(y)/(1-F(t_i^o-)), \tag{2.6}$$

where $F(u|v)$ is defined in (1.5); see Eq. (2.25) of Lai and Ying (1994). Although $F(y-) = F(y)$ since F is continuous, we still write $F(t_i^o-)$ in (2.6), where F will be replaced later by the product-limit estimate which is discrete. In (2.6), it is assumed that F is known; in fact, $E\{h_F(\tilde{y}_i^o, \delta_i^o, t_i^o)\} = (\int_a^b h dF)/P\{y_1 \wedge c_1 \ge t_1\}$, cf. Lemma 1 of Gross and Lai (1996). When F is unknown, we replace it in h_F by the product-limit estimate $\hat{F}^{(\nu)}$ based on the test sample $\mathcal{S}_\nu$ when n/m is not too small, or by the product-limit estimate $\hat{F}$ based on entire sample otherwise. Let $\hat{h}^{(\nu)} = h_{\hat{F}^{(\nu)}}$, $\hat{h} = h_{\hat{F}}$, let $\#(\mathcal{S}_\nu)$ denote the size of the subsample $\mathcal{S}_\nu$. Setting first $h = A_k$ and then $h = D_k$, an alternative to $\sum_{\nu=1}^m L(k, G_\nu, \hat{F}^{(\nu)})$ as the criterion for m-fold cross validation is

$$C_m(k) = \sum_{\nu=1}^m \{ \sum_{(\tilde{y}_i^o, \delta_i^o, t_i^o) \in \mathcal{S}_\nu} a_k^T(G_\nu)\hat{A}_k(\tilde{y}_i^o, \delta_i^o, t_i^o)a_k(G_\nu) + 2a_k^T(G_\nu)\hat{D}_k(\tilde{y}_i^o, \delta_i^o, t_i^o)\}/\#(\mathcal{S}_\nu). \tag{2.7}$$

In the censored case without truncation variables, if we replace $\hat{A}_k$ and $\hat{D}_k$ in (2.7) by $\hat{A}_k^{(\nu)}$ and $\hat{D}_k^{(\nu)}$, then (2.7) reduces to $\sum_{\nu=1}^m L(k, G_\nu, \hat{F}^{(\nu)})$ as a consequence of the following identity due to Susarla, Tsai and Van Ryzin (1984):

$$\sum_{(y_j, \delta_j) \in \mathcal{S}_\nu} \hat{h}^{(\nu)}(y_j, \delta_j)/\#(\mathcal{S}_\nu) = \int h(y)d\hat{F}^{(\nu)}(y). \tag{2.8}$$

In practice, taking $p = 0.05$ and $p^* = 0.95$, which amounts to omitting 5% of either tail of $\hat{F}$, suffices to provide an adequate range of y's at which the efficient score function can be approximated by splines for use in adaptive estimation problems, while maintaining stability of the approximation.

2.4. Numerical examples

Figure 2(a)–(d) represent the true and the estimated score functions based on a simulated dataset of 200 observations $(\tilde{y}_i^o, \delta_i^o, t_i^o)$ generated from each of the following models. The vertical line segments along the horizontal axis represent the uncensored observations.

(a) Lognormal: $\log(y_j) \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $\log(t_j) \overset{\text{i.i.d.}}{\sim} N(-1,1)$, $\log(c_j) \overset{\text{i.i.d.}}{\sim} N[1,1]$,

(b) Contaminated normal: $y_j \overset{\text{i.i.d.}}{\sim} 0.9N(0,1/9)+0.1N(0,9)$ and $t_j \overset{\text{i.i.d.}}{\sim} N(-1,1)$, $c_j \overset{\text{i.i.d.}}{\sim} N(0.8,1)$,

(c) Normal: $y_j \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $t_j \overset{\text{i.i.d.}}{\sim} N(-1,1)$, $c_j = t_j + u_j \cdot \max\{0.5, e^{-t_j}\}$ with $u_j \overset{\text{i.i.d.}}{\sim} U[0,0.1]$,

(d) Beta: $y_j \overset{\text{i.i.d.}}{\sim} Beta(2,2)$ and $t_j \overset{\text{i.i.d.}}{\sim} U[-1,1]$, $c_j \overset{\text{i.i.d.}}{\sim} U[0.5,1]$.

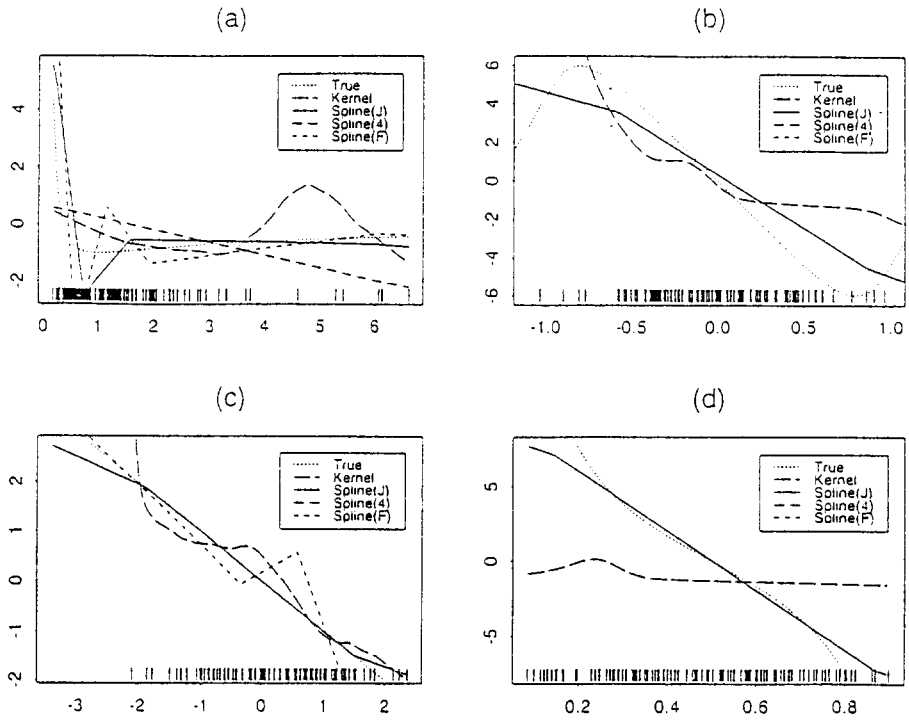

Figure 2: (a) Top left: lognormal, (b) top right: contaminated normal, (c) bottom left: normal, (d) bottom right: beta models.

There are four estimated score functions in each plot. They are labeled Kernel, Spline(J), Spline(4) and Spline(F) respectively. "Kernel" refers to the estimate obtained by smoothing $\hat{\lambda}^{(1)}/\hat{\lambda} - \hat{\lambda}$ using Friedman's supersmoother. The "Spline(·)" estimates refer to the estimated linear spline approximations using different methods to choose the number of knots: Spline(J) uses the extension of Jin's method in Section 2.2; Spline(4) uses the 4-fold cross validation criterion $\sum_{\nu=1}^{4} L(k, G^{(\nu)}, F^{(\nu)})$; Spline(F) uses the full cross validation criterion $C_n(k)$.

In Figure 2(b) and (d), all spline estimates choose the same number of knots and therefore agree with each other because of the way (2.4) in which knots are placed. In Figure 2(c), Jin's method and 4-fold cross validation pick no knot between a and b while full cross validation chooses 2 knots. In Figure 2(a)-(c), all estimates are quite close to the true score function. However, in Figure 2(d), the kernel estimate is relatively flat and differs substantially from the true score function, which is well approximated by the spline estimates that coincide with each other.

Table 1: Comparison of the mean squared errors (MSE) of different estimates of the efficient score function in four models, whose censoring proportion p_c and truncation proportion p_t are also indicated.

Model	**Estimate**	**MSE**	**SE**
Normal	Kernel	0.151	0.096
$p_c = 0.27$	Spline(J)	0.074	0.023
$p_t = 0.28$	Spline(4)	0.075	0.022
	Spline(F)	0.050	0.024
Contaminated	Kernel	3.001	1.100
normal	Spline(J)	1.499	0.792
$p_c = 0.26$	Spline(4)	1.220	1.030
$p_t = 0.28$	Spline(F)	0.580	0.320
Lognormal	Kernel	4.123	2.711
$p_c = 0.27$	Spline(J)	1.401	0.542
$p_t = 0.26$	Spline(4)	1.799	0.699
	Spline(F)	1.489	0.811
Beta	Kernel	18.765	10.110
$p_c = 0.28$	Spline(J)	7.234	7.310
$p_t = 0.29$	Spline(4)	8.792	7.011
	Spline(F)	4.516	2.342

Table 1 compares the mean squared errors

$$\text{MSE} = E(\hat{\psi}(Y) - \psi^*(Y))^2$$

of the estimates $\hat{\psi}$ obtained from $(\tilde{y}_i^o, \delta_i^o, t_i^o)$, $i = 1, \cdots, 200$, by these four methods, where Y is generated from F and is independent of the (y_j, c_j, t_j) underlying the observed $(\tilde{y}_i^o, \delta_i^o, t_i^o)$. Each MSE in Table 1 is based on 100 simulations, and its associated standard error (SE) is also included in the table. The results in Table 1 show that the kernel estimate has considerably larger MSE than the spline estimates, and that the spline estimate with full cross validation tends to have a smaller MSE than the other spline estimates.

3. CONCLUSION

For regression analysis with complete data, the least square estimate is widely used because of its simplicity that may have inferior performance if the errors are non-normal. Using a nonlinear score function that differs from $\psi(x) = x$ for the least squares estimate leads to an M-estimator with greater computational complexity but with better robustness properties. For l.t.r.c. data, there are no computational advantages in choosing $\psi(x) = x$ for the estimating equation (1.2) defining M-estimators. In general proper choice of ψ depends on the underlying distribution F of the ϵ_j. Our numerical study shows that for samples of size 100 and larger, one can estimate ψ reasonably well and achieve good performance of the adaptive estimator of the score function by using regression splines and a relatively simple cross validation method for determining the number of knots, despite substantial truncation and censoring of the response variable.

REFERENCES

Amemiya, T. (1985). *Advanced Econometrics*, Harvard Univ. Press, Cambridge.

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

Bickel, P.J. (1982). On adaptive estimation. *Annals of Statistics*, **10**, 647–671.

Faraway, J.J. (1992). Smoothing in adaptive estimation. *Annals of Statistics*, **20**, 414–427.

Goldberger, A.S. (1981). Linear regression after selection. *Journal of Econometrics*, **15**, 357–366.

Gross, S. and Lai, T.L. (1996). Nonparametric estimation and regression analysis with left truncation and right censored data. *Journal of the American Statistical Association*, **91**, 1166–1180.

Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge Univ. Press.

Hsieh, D.A. and Manski, C.F. (1987). Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. *Annals of Statistics*, **15**, 541–551.

Jin, K. (1992) Empirical smoothing parameter selection in adaptive estimation. *Annals of Statistics*, **20**, 1844–1874.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

Keiding, N., Holst, C. and Green, A. (1989). Retrospective estimation of diabetes incidence from information in a current prevalent population and historical mortality. *American Journal of Epidemiology*, **130**, 588–600.

Kim, C.K. and Lai, T.L. (1999). Robust regression with censored and truncated data. In *Multivariate Analysis, Design of Experiments and Survey Sampling* (S. Ghosh, ed.), Marcel Dekker, New York.

Lai, T.L. and Ying, Z. (1991a). Estimating a distribution function with truncated and censored data. *Annals of Statistics*, **19**, 417–442.

Lai, T.L. and Ying, Z. (1991b). Rank regression methods for left-truncated and right-censored data. *Annals of Statistics*, **19**, 531–556.

Lai, T.L. and Ying, Z. (1992). Asymptotically efficient estimation in censored and truncated regression models. *Statistica Sinica*, **2**, 17–46.

Lai, T.L. and Ying, Z. (1994). A missing information principle and M-estimators in regression analysis with censored and truncated data. *Annals of Statistics*, **22**, 1222–1255.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

Lin, D.Y. and Geyer, C.J. (1992). Computational methods for semiparametric linear regression with censored data. *Journal of Computational and Graphical Statistics*, **1**, 77–90.

Moon, C-G. (1989). A Monte Carlo comparison of semiparametric Tobit estimators. *Journal of Applied Econometrics*, **4**, 361–382.

Nicoll, J.F. and Segal, I.E. (1980). Nonparametric estimation of the observational cutoff bias. *Astron. Astrophy.*, **82**, L3–L6.

Segal, I.E. (1975). Observational validation of the chronometric cosmology. I. Preliminaries and the redshift magnitude relation. *Proceedings of National Academic Sciences, USA*, **72**, 2437–2477.

Susarla, V. Tsai, W.Y. and Van Ryzin, J. (1984). A Buckley-James-type estimator for the mean with censored data. *Biometrika*, **71**, 624–625.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24–36.

Uzunoğullari, Ü and Wang, J.L. (1992). A comparison of hazard rate estimators for left-truncated and right-censored data. *Biometrika*, **79**, 297-310.