▩ 연구논문

중도절단된 자료에 대한 가법회귀모형

김철기
이화여대 통계학과

Additive Regression Models for Censored Data

Chul-Ki Kim
Dept. of Statistics, Ewha Womans University

Abstract

In this paper we develop nonparametric methods for regression analysis when the response variable is subject to censoring that arises naturally in quality engineering. This development is based on a general missing information principle that enables us to apply, via an iterative scheme, nonparametric regression techniques for complete data to iteratively reconstructed data from a given sample with censored observations. In particular, additive regression models are extended to right-censored data. This nonparametric regression method is applied to a simulated data set and the estimated smooth functions provide insights into the relationship between failure time and explanatory variables in the data.

Key words: Additive regression models; Right-censored data; Missing information principle; Smoothing.

1. Introduction

Within the past two decades, semiparametric methods have been developed for regression analysis when the response variable Y is subject to right censoring. One approach in this semiparametric regression modelling is to use Cox's proportional hazard model which assumes that the hazard function of Y when covariate has the value x has the form

$$\lambda(y \mid x) = \lambda_0(y)\exp(-\beta^T x) \tag{1.1}$$

in which λ_0 is an unknown baseline hazard function (at $x=0$) and β is the vector of unknown regression coefficients, cf. Cox (1975). Another approach is to assume the standard regression model

$$y = \alpha + \beta^T x + \varepsilon \tag{1.2}$$

in which ε represents an unobservable random disturbance, independent of x, that has an unknown distribution function F with mean zero, as in Miller (1976), Buckley and James (1979), Susarla and Van Ryzin (1984), Ritov (1990), and Lai and Ying (1994), among others.

In this paper, we consider the following generalization of (1.2):

$$y_i = g(x_i) + \varepsilon_i \ (i=1, \cdots, n) \tag{1.3}$$

in which the common distribution F of the independent and identically distributed (i.i.d.) ε_i is unknown. The y_i in (1.2), however, are not completely observable due to right censoring by $c_i(\leq \infty)$. Set $c_i \equiv \infty$ if there is no censoring. An advantage of (1.3) over (1.2) is the flexible functional form of the relationship between the predictors and response.

We show how nonparametric regression methods which have been found to work well for complete data can be extended to right-censored data.

By imposing some additive structure on the regression function g, addiitive regression models (Hastie and Tibshirani, 1990) have proved to be effective in circumventing the curse of dimensionality in high dimensions of x_i in (1.3). We extend additive regression models (ARM) in Section 2 to right-censored data. Part of the success of ARM stems from the availability of efficient algorithms to implement these computer-intensive nonparametric regression methods. In Section 3 we show how this algorithm can be augmented and modified for right-censored data and we use it in a simulation example in Section 4.

2. Additive Regression Models for Complete Data

An additive regression model is defined by

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon \tag{2.1}$$

where the errors ε are independent of the X_j, $E(\varepsilon)=0$ and $Var(\varepsilon)=\sigma^2$. The f_j, one for each predictor, are arbitrary univariate functions. Implicit in (2.1) is the assumption that $E\{f_j(X_j)\}=0$, since there would otherwise be free constants in each of the functions. All the f_j are assumed to be smooth functions, and they are somehow individually estimated by an arbitrary smoother.

The *backfitting algorithm* is a general algorithm that enables one to fit an additive regression model using any regression-type fitting mechanisms (cf. Hastie and Tibshirani, 1990). It is an iterative-fitting procedure. Conditional expectations provide a simple intuitive motivation for the backfitting algorithm. If the additive model (2.1) is correct, then for any k, $E(Y-\alpha-\sum_{j\neq k} f_j(X_j)\mid X_k) = f_k(X_k)$, which corresponds to finding the minimizers of $E[(Y-\alpha-\sum_{j=1}^{p} f_j(X_j))^2]$. This immediately suggests for computing all the f_j an iterative algorithm, which is given in terms of data and arbitrary scatterplot smoothers (cf. Buja, Hastie and Tibshirani, 1989).

The Backfitting Algorithm

Initialize $\mathbf{a} = \text{ave}(\mathbf{y})$, $\mathbf{f}_j = \mathbf{f}_j^0$; $j=1, \cdots, p$

Cycle $j = 1, \cdots, p, 1, \cdots, p, \cdots$

$$\mathbf{f}_j = S_j[\mathbf{y}-\mathbf{a}-\sum_{k\neq j}\mathbf{f}_k \mid \mathbf{x}_j]$$

Continue cycle until the individual functions do not change.

Here $S_j[\mathbf{y}\mid\mathbf{x}_j]$ denotes a smooth of the response $\mathbf{y}$ against the predictor $\mathbf{x}_j$, and produces a function. We want the functions to be fitted simultaneously, so the individual smoothing steps make sense. When readjusting $\mathbf{f}_j$, we remove the effects of all the other variables from $\mathbf{y}$ before smoothing this *partial residual* against $\mathbf{x}_j$. Clearly this is only appropriate if all the functions removed are also correct, and therefore, iteration is involved.

We need to provide initial functions to start the algorithm. Without prior knowledge of the functions, a sensible starting point might be the linear regression of $\mathbf{y}$ on the predictors. Often the backfitting algorithm itself is nested within some bigger iteration, in which case the functions from the previous big iteration loop provide starting values (cf. Section 3).

3. Additive Regression Models for Right-Censored Data

Already described in Section 2, the minimization problem in nonparametric additive regression for complete data is the following

· ARM (Additive Regression Model):

$$\sum_{i=1}^{n}(y-\alpha-\sum_{j=1}^{p}f_j(x_{ji}))^2 = \min! \tag{3.1}$$

How can this optimization criterion be extended when the (x_i, y_i) are not completely observable because of right censoring on the y_i? To answer this question, we shall use a generalization of the missing information principle of Lai and Ying (1994) to nonparametric regression models.

The basic underlying idea in this missing information principle is to iterate the following two steps until the entire procedure converges: the E-step consists of estimating the conditional expectations of nonparametric normal equations based on the observed right-censored data; and the M-step consists of solving the normal equations associated with the optimization problem. The conditional expectations involve the unknown distribution function F of the errors ε_i and thus we have to also estimate F for the E-step.

First, consider the minimization problem (3.1) in the additive regression model

$$y_i = \alpha + \sum_{j=1}^{p} f_j(x_{ji}) + \varepsilon_i \quad (i=1, \cdots, n), \tag{3.2}$$

where the ε_i are i.i.d. random variables with a common continuous distribution function F having a finite mean (not necessarily 0) and without loss of generality each underlying function in (3.2) is constrained to sum to zero.

Suppose that the responses y_i in (3.2) are not completely observable due to right censoring by random variable c_i such that $-\infty < c_i \leq \infty$. Let $\tilde{y}_i = y_i \wedge c_i$ and $\delta_i = I(y_i \leq c_i)$, where we use $\wedge$ to denote minimum. The data, therefore, consist of n observations $(x_i, \tilde{y}_i, \delta_i)$, $i=1, \cdots, n$. Using this notation, we naturally modify the criterion by taking the conditional expectation to (3.1) based on the right-censored data

$$\sum_{i=1}^{n} E(y_i-a-\sum_{j=1}^{p} f_{ji})^2 \mid \tilde{y}_i, \delta_i] = \min!, \tag{3.3}$$

where f_{ji} denotes $f_j(x_{ji})$. We can express (3.3) in vector form as

$$E[(\mathbf{y}-\mathbf{a}-\sum_{j=1}^{p}\mathbf{f}_j)^T(\mathbf{y}-\mathbf{a}-\sum_{j=1}^{p}\mathbf{f}_j)\mid\tilde{\mathbf{y}},\delta]=\min!, \tag{3.4}$$

where $\mathbf{y}=(y_1,\cdots,y_n)^T$, $\mathbf{a}=(1,\cdots,1)^T$, $\mathbf{f}_j=(f_{j1},\cdots,f_{jn})^T$, $\tilde{\mathbf{y}}=(\tilde{y}_1,\cdots,\tilde{y}_n)^T$, and $\delta=(\delta_1,\cdots,\delta_n)^T$.

Let $\hat{F}_{n,f}$ denote the product-limit curve based on $(\tilde{y}_i(\mathbf{f}),\delta_i)_{i\le n}$, where $\tilde{y}_i(\mathbf{f})=\tilde{y}_i-\sum_{j=1}^{p}f_{ji}$. For $i=1,\cdots,n$, let

$$Z_n(\mathbf{f},u)=\sum_{i=1}^{n}I(\tilde{y}_i(\mathbf{f})\ge u),\ \Delta(\mathbf{f},u)=\sum_{i=1}^{n}I(\tilde{y}_i(\mathbf{f})=u,\ \delta_i=1) \tag{3.5}$$

and define

$$\hat{F}_{n,f}(u)=1-\prod_{i:\tilde{y}_i(f)\le u,\ \delta_i=1}(1-\Delta(\mathbf{f},u)/Z_n(\mathbf{f},\tilde{y}_i(\mathbf{f}))). \tag{3.6}$$

Since the conditional expectation in (3.3) involves the unknown common distribution function F, it is natural to replace F by $\hat{F}_{n,f}$ in (3.6). The unrealizable equation (3.3), even with F replaced by $\hat{F}_{n,f}$, suggests the following iterative mechanism, which is indeed an analog of the EM algorithm for ARM under right censorship:

· Substitute $\mathbf{f}_j$ and F by $\hat{\mathbf{f}}_j^{(k-1)}$ and $\hat{F}_{n,\hat{f}^{(k-1)}}$. This gives

$\hat{E}^{(k)}[(\mathbf{y}-\mathbf{a}-\sum_{j=1}^{p}\mathbf{f}_j)^T(\mathbf{y}-\mathbf{a}-\sum_{j=1}^{p}\mathbf{f}_j)\mid\tilde{\mathbf{y}},\delta]$. Note that for the right censored data $(x_i,\tilde{y}_i,\delta_i)$

$$E(y_i\mid\tilde{y}_i,\delta_i)=\tilde{y}_i+(1-\delta_i)\int_{\tilde{y}_i(f)}^{\infty}(1-F(v))dv/(1-F(\tilde{y}_i(\mathbf{f}))) \tag{3.7}$$

and the conditional expectation in the kth iteration can be evaluated by replacing $\mathbf{f}_j$ and F by $\hat{\mathbf{f}}_j^{(k-1)}$ and $\hat{F}_{n,\hat{f}^{(k-1)}}$.

· Define $\hat{\mathbf{a}}^{(k)}$ and $\hat{\mathbf{f}}_j^{(k)}$ as a solution of the normal equation

$$[\hat{E}^{(k)}(\mathbf{y}\mid\tilde{\mathbf{y}},\delta)-\mathbf{a}-\sum_{j=1}^{p}\mathbf{f}_j]=\mathbf{0}, \tag{3.8}$$

which corresponds to minimizing

$$\hat{E}^{(k)}[(\mathbf{y}-\mathbf{a}-\sum_{j=1}^{p}\mathbf{f}_j)^T(\mathbf{y}-\mathbf{a}-\sum_{j=1}^{p}\mathbf{f}_j)\mid\tilde{\mathbf{y}},\delta]. \tag{3.9}$$

Here $\hat{\mathbf{f}}_j^{(k)}$, $\hat{\mathbf{a}}^{(k)}$, $\hat{F}_n^{(k)}$ and $\hat{E}^{(k)}[\cdot\mid\tilde{\mathbf{y}},\delta]$ denote the kth evaluations at the n observations $(x_i,\ \tilde{y}_i,\ \delta_i)$.

To implement the nonparametric estimating equation (3.8) we first recall the *backfitting algorithm* of ARM (See Section 2). In the algorithm $S_j[\cdot\mid\mathbf{x}_j]$ stands for smoothing the *partial residuals* on $\mathbf{x}_j$. Compared with the estimating equation in the backfitting algorithm, the normal equation (3.8) equivalent to the following estimating equation

$$\hat{\mathbf{f}}_j=\mathrm{S}_j[\hat{E}^{(k)}(\mathbf{y}\mid\tilde{\mathbf{y}},\delta)-\mathbf{a}-\sum_{k\neq j}\mathbf{f}_k\mid\mathbf{x}_j]. \tag{3.10}$$

Now we shall introduce the ARM-RC (Additive Regression Model under Right Censorship) algorithm which is a two-stage backfitting algorithm deduced from (3.10): first, given the estimated conditional expectations of dependent variable and the underlying functions, do the authentic backfitting with a smoother; and at the second stage, use the fitted functions to compute the conditional expectations. Iterate these loops until each one converges and the convergence criterion is $\|V_{new}-V_{old}\|/\|V_{old}\|\leq threshold$ with $\|a\|=\sqrt{a^Ta}$. In applications, we use 0.00003 for the threshold and a variable span smoother (the so-called *supersmoother* of Friedman (1984)) is recommended for the smoothing procedure in backfitting. The term "*Smooth*" in the ARM-RC algorithm refers to the supersmoother. This smoother is a nonlinear smoother and an enhancement of the running-line smoother, the difference being that it chooses a (possibly) different span at each X-value. It does so in order to adapt the changes (across X) in the curvature of the underlying function and the variance of Y. In regions where the curvature-to-variance ratio is high, a small span is appropriate, while in low curvature-to-variance regions a large span is called for. The supersmoother tries to achieve this effect as follows. Three windows of small, medium and large spans are passed over the data. For each span, the squared cross-validated residual is computed at each point and smoothed as a function of X. Then the span producing the smallest smoothed squared-residual at $X=x_i$ is chosen. Finally, the optimal span values are smoothed against X so that the spans do not change too quickly as X varies. In simulations, the supersmoother seems able to adapt the span appropriately although some price is paid in increased variance of the estimate. Despite its complex nature, there is an $O(n)$ algorithm for supersmoother that makes repeated

use of the updating formula (cf. Friedman, 1984).

The ARM-RC Algorithm

Initialize

$\mathbf{y}^* \leftarrow$ synthesize $\tilde{\mathbf{y}}$ if censored

$\hat{f}_j(\mathbf{x}_j) \leftarrow f_j^0; \quad j=1, \cdots, p$

Loop 1

$\hat{\alpha} \leftarrow \text{Ave}(\mathbf{y}^*)$

Loop 2 $j=1, \cdots, p$

$$\hat{f}_j(\mathbf{x}_j) \leftarrow Smooth(\mathbf{y}^* - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(\mathbf{x}_k) \mid \mathbf{x}_j)$$

Do Loop 2 Until change in $\{\hat{f}_j(\mathbf{x}_j)\}_{j \le p} < threshold$

Compute the Product-Limit estimator based on $(\hat{\mathbf{y}} - \sum_{j=1}^{p} \hat{f}_j(\mathbf{x}_j), \delta)$

Do Loop 1 Until change in $\mathbf{y}^*$'s $<$ *threshold*.

In Loop 1 of the ARM-RC algorithm above, the $\mathbf{y}^*$ represents a vector of the n conditional expectations (3.7) evaluated in each iteration (E-step). Once the $\mathbf{y}^*$ is computed, Loop 2 does the same job as the backfitting algorithm for complete data (M-step). That is, Loop 2 solves the estimating equation (3.10). After Loop 2 updates the smooths, Loop 1 computes the product-limit curve based on the estimated smooths and synthesized residuals to update the $\mathbf{y}^*$. The entire procedure iterates until the changes in $\mathbf{y}^*$ and $\hat{f}_j\{(\mathbf{x}_j)\}_{j \le p} <$ *threshold*. In applications, the default on the maximum number of iterations is set at 200.

4. Numerical Examples

The following simulation studies show how well this algorithm works.

Example. Simulated censored data

Two hundred data points are generated from the following simulation model:

$$y_i = x_{1i}^2 + \sin(\pi \cdot x_{2i}) + \varepsilon_i \quad (i=1, \cdots, 200) \tag{4.1}$$

where

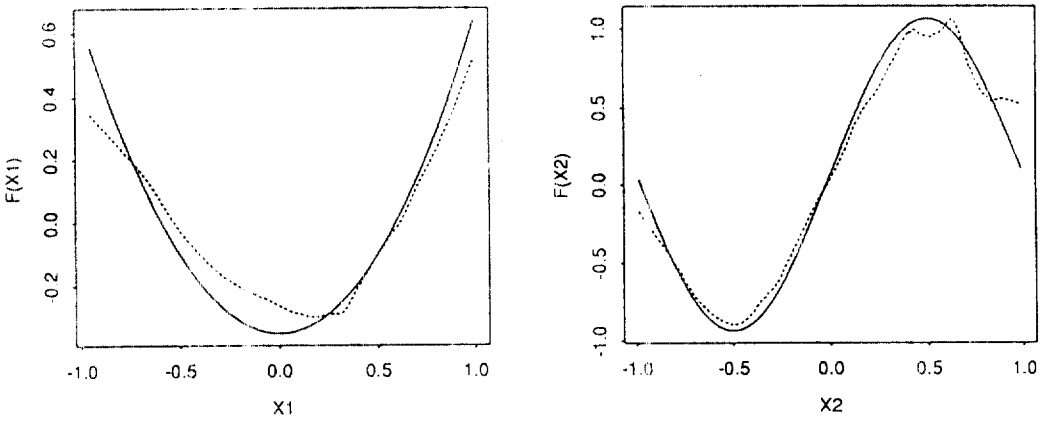
$\varepsilon_i \sim iid\ N(0.0\,4^2)$

x_{i1} and $x_{i2} \sim iid\ Uniform\ [-1, 1]$,

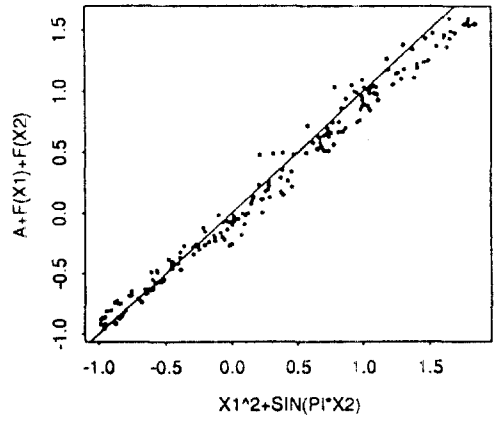
$c_i \sim iid\ Uniform\ [-3, 5]$ with probability 1/2 and Centered Extreme Value Distribution with probability 1/2, and

42% of data points are censored.

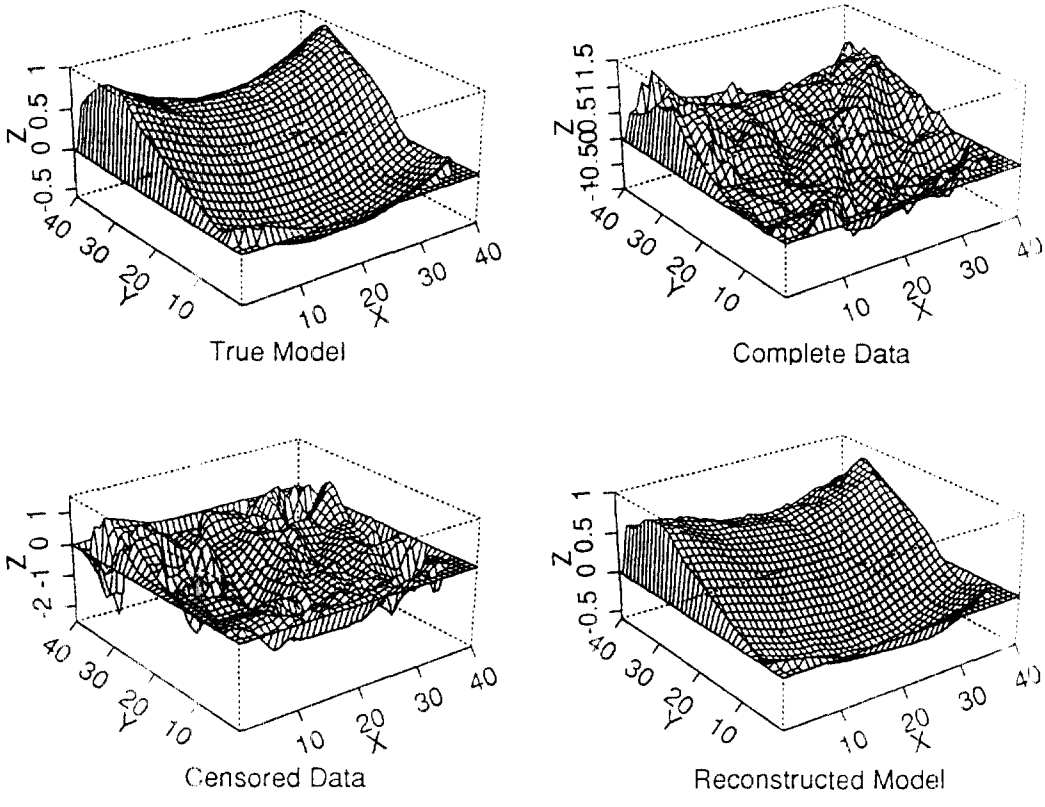
Herein, note that the extreme value distribution is defined as $1-F(t)=\exp(-\exp(t))$. The two plots in 〈Figure 1〉 show how well $\hat{f}_1(\mathbf{x}_1)$ and $\hat{f}_2(\mathbf{x}_2)$ are estimated by the algorithm. As defined in simulation model (4.1), the true $f_1(\mathbf{x}_1)$ and $f_2(\mathbf{x}_2)$ are a square and a sine curve respectively. As we see in the panels, the estimated curves are quite close to the true ones. 〈Figure 2〉 also shows that the estimated regression curve fits the model well. Moreover, the four 3-D plots provided in 〈Figure 3〉 help us visualize how the procedure works more clearly. Note that the numeric values and shape of the reconstructed model are almost the same as those of the true model (cf. Figure 3). The two 3-D plots in 〈Figure 4〉 compare an average shape of 100 replications to the true model. Due to the boundary effect, the bands that are most to the right and most to the left in 〈Figure 5〉 are, as expected, wider than any other areas. And the second panel shows that the supersmoother somewhat oversmoothes the sine curve as its three predetermined spans---0.05, 0.2, 0.5---are all larger than the appropriate window size for the curvature present in the true function. The histogram 〈Figure 6〉 tells us that for most of the 100 replications from the simulation model (4.1) the loop converges after 7 or 8 iterations on the average.



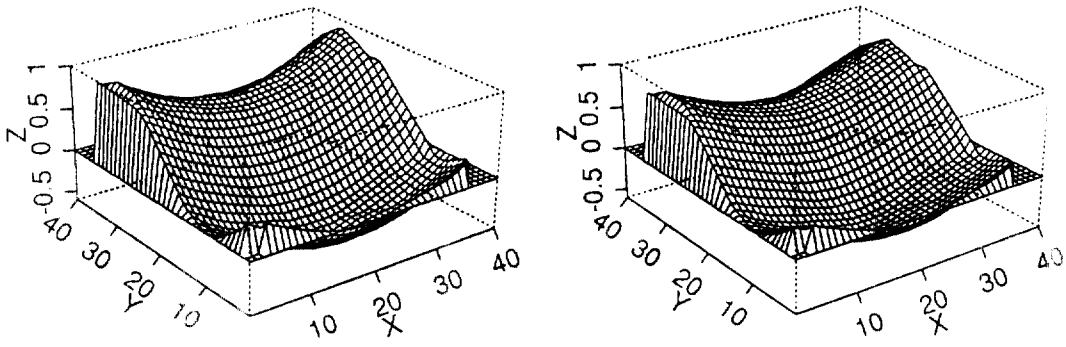
〈 Figure 1 〉 The true and the estimated functions for ARM with censored data represented by the solid and the dotted lines respectively.



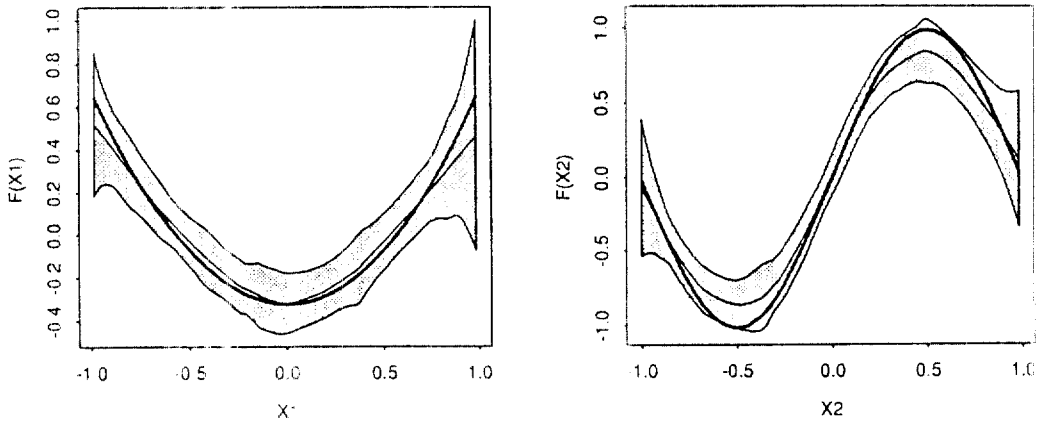
〈 Figure 2 〉 Plot of the simulated data versus the fitted plot



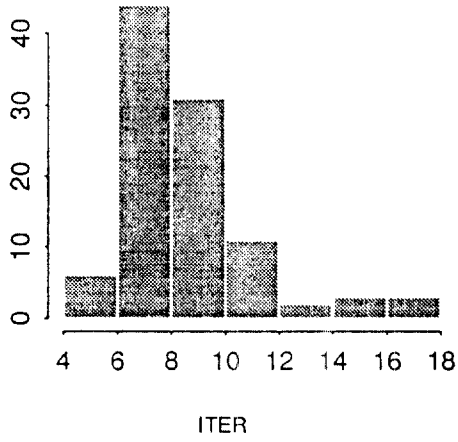
〈 Figure 3 〉 Actual model, data plots and reconstructed model for ARM with censored data X, Y and Z represents x_1, x_2 and $\hat{y}$.



〈 Figure 4 〉 The true model versus that averaged from 100 replications.



〈 Figure 5 〉 Average curves of 100 replications for each function in ARM with censored data. The thick and thin solid lines represent the true functions and the estimated smooths, respectively. The light shaded regions show twice the pointwise standard errors of the fitted smooth values and, thus, reflect the variance of the smooths under repeated sampling of 200 4-tuples from model (4.1).



〈 Figure 6 〉 Distribution of the number (ITER) of iterations in repeated sampling of 100 times.

5. Conclusion

Most of parametric regression models for failure time data can be transformed to "the accelerated failure time model" (cf. Kalbfleisch and Prentice, 1980) by taking logarithm to the failure time. This log-linear model leads to the problem of estimating the regression function that is linear in the parameter β and it is the precursor to the Cox proportinal hazard model making use of partial (conditional) likelihood. Moreover, as we mentioned in Section 1, using the general missing information principle (cf. Lai and Ying, 1994), we can generalize the regression models (1.2) and (1.3) to the nonparametric regression problem between failure time and covariates under censorship, which helps us to find the true functional relationship between them. Thus this nonparametric regression technique hopefully enables us to control failure time under censorship more effectively based on covariates.

References

[1] Buckley, J. and James, I. (1979), "Linear regression with censored data," *Biometrika*, Vol. 66, pp. 429-436.

[2] Buja, A., Hastie, T. and Tibshirani, R. (1989), "Linear smoothers and additive models (with discussion)," *Ann. Statist.*, Vol. 17, pp. 453-555.

[3] Cox, D. R. (1975), "Partial likelihood," *Biometrika*, Vol. 62, pp. 269-276.

[4] Dempster, A. P., Laired, N. M. and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *J. Roy. Statist. Soc. Ser. B*, Vol. 39 pp. 1-38.

[5] Efron, B. (1967), "The two sample problem with censored data," *Proc. Fifth Berkeley Symp. Math. Statist. Probab.*, Vol. 4, pp. 831-853.

[6] Friedman, J. H. (1984), *A variable span smoother*, Stanford University, Department of Statistics Technical Report LCS5.

[7] Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall, New York.

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

[9] Lai, T. L. and Ying, Z (1994), "A missing information principle and M-estimator in regression analysis with censored and truncated data," *Ann. Statist.*, Vol. 22, pp. 1222-1255.

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

[11] Miller, R. G. (1976), "Least squares regression with censored data," *Biometrika*, Vol. 63, pp. 449-464.

[12] Miller, R. G. and Halpern, J. (1982), "Regression with censored data," *Biometrika*, Vol. 69, pp. 521-531.

[13] Orchard, T. and Woodbury, M. A. (1972), "A missing information principle: theory and applications," *Proc. Sixth Berkeley Symp. Math. Statist. Probab.*, Vol. 1, pp. 697-715.

[14] Ritov, Y. (1990), "Estimation in a linear regression model with censored data," *Ann. Statist.*, Vol. 18, pp. 303-328.

[15] Susarla, V., Tsai, W. Y. and Van Ryzin, J. (1984), "A Buckeley-James-type estimator for the mean with censored data," *Biometrika*, Vol. 71, pp. 624-625.