Universidade de Vigo

Additive Models in Censored Regression

Jacobo de Uña Álvarez and Javier Roca Pardiñas.

Report 08/01

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://eioweb.uvigo.es/
E-mail: depc05@uvigo.es

Universidade de Vigo

Additive Models in Censored Regression

Jacobo de Uña Álvarez and Javier Roca Pardiñas.

Report 08/01

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita: Universidade de Vigo
Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

Additive Models in Censored Regression

Jacobo de Uña Álvarez
Departamento de Estadística e I.O.
Universidad de Vigo

Javier Roca Pardiñas
Departamento de Estadística e I.O.
Universidad de Vigo
c/Torrecedeira 86
E-36208 Vigo

Abstract

In this paper we consider additive models in censored regression. We propose a randomly weighted version of the backfitting algorithm that allows for the nonparametric estimation of the effects of the covariates on the response. Given the high computational cost involved, binning techniques are used to speed up the computation in the estimation and testing process. Simulation results and the application to real data reveal that the predictor obtained with the additive model performs well, and that it is a convenient alternative to the linear predictor when some nonlinear effects are suspected. [1]

Keywords: additive models, backfitting

1 Introduction

Let Y be a lifetime which is observed under censoring from the right. Let $\mathbf{X} = (X_1, ..., X_p)'$ be a vector of p covariates. Put $f(\mathbf{x}) = E\left[\psi(Y) \mid \mathbf{X} = \mathbf{x}\right]$

[1] Javier Roca Pardiñas was supported by grants MTM2005-00818 (European FEDER funding included) and SEJ2004-04583 /ECON. Jacobo de Uña Álvarez was supported by grants MTM2005-01274 (European FEDER funding included) and PGIDIT06PXIC300117PN.

for the regression function of $\psi(Y)$ on $\mathbf{X}$, so the model becomes

$$\psi(Y) = f(\mathbf{X}) + \varepsilon = f(X_1, ..., X_p) + \varepsilon \tag{1}$$

where the error term satisfies $E[\varepsilon \mid \mathbf{X}] = 0$. Here ψ denotes a time transformation such as the logarithm. Taking $\psi(y) = \ln y$ is useful in regression analysis because $\psi(Y)$ is no longer restricted to $(0, \infty)$. Indeed, under (1) we have, provided that $\mathbf{X}$ and ε are independent,

$$F_{Y|\mathbf{X}}(y \mid \mathbf{X}) = F_W(e^{-f(\mathbf{X})}y), \qquad y \geq 0,$$

where $F_{Y|\mathbf{X}}$ and F_W are the cumulative distribution functions of Y given $\mathbf{X}$ and of the transformed error $W = e^{\varepsilon}$, respectively. This is the so-called accelerated failure time model, widely used to analyze survival data in the regression framework. Note that an increasing value of $f(\mathbf{X})$ results in a decreasing value of the time acceleration factor $e^{-f(\mathbf{X})}$, thus leading to a better survival prognosis.

In the censored setup, we observe $(\mathbf{X}_1, Z_1, \delta_1), ..., (\mathbf{X}_n, Z_n, \delta_n)$ independent observations with the same distribution as $(\mathbf{X}, Z, \delta)$, where $Z = \min(Y, C)$, C is the right-censoring variable assumed to be independent of Y, and $\delta = \mathbb{I}(Y \leq C)$. Unlike in the "iid" scenario (in which each observation receives mass or weight $1/n$), the weight associated to the i-th observation $(\mathbf{X}_i, Z_i, \delta_i)$ under censoring will be typically the jump of the Kaplan-Meier estimator at each point Z_i $(i = 1, ..., n)$, namely

$$W_i = \frac{\delta_i}{n - RankZ_i + 1} \prod_{RankZ_j < RankZ_i} \left[1 - \frac{\delta_j}{n - RankZ_j + 1} \right]$$

where $RankZ_i$ is the rank of Z_i among the ordered Z's and where (in case of ties) uncensored observations are assumed to preceed the censored ones. When the error distribution is unknown, an approach that leads to consistent estimators is choosing f in order to minimize

$$f \in \mathcal{F} \mapsto \sum_{i=1}^{n} W_i \left(\psi(Z_i) - f(X_{i1}, ..., X_{ip}) \right)^2$$

where the family $\mathcal{F}$ represents the *a priori* knowledge on the true regression. See Stute (1993, 1996, 1999) for the parametric linear and nonlinear case, in which it is assumed $f \in \{f(.;\beta)\}_\beta$, and see Orbe, Ferreira and Núñez-Antón (2003) for the partly linear case $f(\mathbf{X}) = \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + g(X_p)$. Another possible approach (which we will not follow here) is that based in the so-called synthetic data, see for example Leurgans (1987) and Qin and Jing (2000) who considered the parametric linear case and the partial linear model, respectively (see also Liang and Zhou, 1998, for the latter setup).

In some instances linear model can be very restrictive. This constraint can be avoided by replacing the linear index with a non-parametric structure. Here we consider a flexible approach to estimate the regression function $f(x)$ through a semiparametric model under which the effect of each covariate on the response is represented in an additive way, the qualitative form of this effect being unknown otherwise. We assume the additive model

$$f(\mathbf{X}) = f_1(X_1) + ... + f_p(X_p) \tag{2}$$

(Hastie and Tibshirani, 1990), where α is a constant and $f_1, ..., f_p$ are one dimensional functions. If the influence of the covariates X_j is linear, then the corresponding partial functions can be expressed parametrically as $f_j(X_j) = \beta_j X_j$. Therefore, the model given in (2) nests the linear model. Moreover, on assuming that effects are additive, this type of models maintain the interpretability of linear models. Yet, at the same time, they incorporate the flexibility of non-parametric smoothing methods because, rather than following a fixed parametric form, the effect of each of the covariates, X_j, depends on a totally unknown function, f_j, which is only required to possess a certain degree of smoothness so that it can be estimated. The compromise between flexibility, dimensionality and interpretability, ranks these types of models among the statistical tools with the greatest capacity for data analysis in different fields of research.

Additive models have been used for relaxing the linear hypothesis in the scope of Cox proportional hazards model, which is the most popular regression model when analyzing censored survival data. For example, Huang (1999) introduced efficient estimation for a partly additive Cox model. Huang and Liu (2006) considered a nonparametric link function which controls for the effect of the parametric predictor under proportional hazards. See also Ganguli and Wand (2006) and references therein for extensions of the Cox proportional hazards model via additive regression. However, the proportional hazards assumption may not hold in some applications, and hence there is some need of additive models which can be a valid alternative to Cox regression. For the best of our knowledge, additive models in the scope of the censored accelerated failure time model have not been explored so far. This is the gap we fill through the present work.

The layout of this paper is as follows. In Section 2 is given a description of weighted kernel smoothing backfitting we use for fitting additive models with censored response. Moreover, in Section 2.1 we discuss the bandwidth selection problem and some related practical issues. To assess the validity of this estimation procedure, a simulation study is performed in Section 3. In Section 4 we apply the proposed methodology to real data. Finally, we conclude with a discussion in Section 5.

2 Fitting censored additive models

This section describes an algorithm for fitting the model effects $f_1, \ldots, f_p$ in (2) for censored response. The algorithm discussed below is a modification of the backfitting algorithm (Opsomer, 2000) used for fitting additive models. The backfitting algorithm cycles through the covariates X_j ($j = 1, \ldots, p$), and estimates each f_j by applying local linear kernel smoothers to the partial residuals. These residuals are obtained by removing the estimated effects of the other covariates. Although our focus is on local scoring, there are other types of procedures that allow for non-parametric estimation of General Additive Models (GAMs). Sperlich *et al.* (2002) presented methods based on marginal integration. Wahba (1990) and Guo (2002) proposed the use of smoothing spline ANOVA methods. Coull *et al.* (2001) and Ruppert *et al.* (2003) investigated alternative methods based on penalised splines, and Wood (2003) used thin plate regression splines. Other studies, such as the paper by Brezger & Lang (2006), also used P-splines, and developed Bayesian versions of GAMs and extensions to generalised structured additive regression.

Given a sample $\{(\mathbf{X}_i, Z_i, \delta_i)\}_{i=1}^n$ of $(\mathbf{X}, Z, \delta)$, the steps of the estimation algorihm are as follows:

Initialisation. Compute the initial estimates $\hat{\alpha} = \sum_{i=1}^n W_i \psi(Z_i) / \sum_{i=1}^n W_i$ and $\hat{f}_j^0 (X_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$

Step 1. For $j = 1, \ldots, p$ calculate residuals by removing the estimated effects of all the other covariates:

$$Z_i^j = \psi(Z_i) - \hat{\alpha} - \sum\nolimits_{s=1}^{j-1} \hat{f}_s (X_{is}) - \sum\nolimits_{s=j+1}^{p} \hat{f}_s^0 (X_{is}),$$

and compute for $i = 1, ..., n$ the weighted local linear kernel estimators (Wand and Jones, 1995)

$$\hat{f}_j (X_{ij}) = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} s_j^0 (X_{ij}) & s_j^1 (X_{ij}) \\ s_j^1 (X_{ij}) & s_j^2 (X_{ij}) \end{pmatrix}^{-1} \begin{pmatrix} u_j^0 (X_{ij}) \\ u_j^1 (X_{ij}) \end{pmatrix} \tag{3}$$

where $s_j^r(x) = \sum_{i=1}^n \left(W_i \cdot L_j^r(x, X_{ij})\right)$ and $u_j^r(x) = \sum_{i=1}^n \left(W_i \cdot L_j^r(x, X_{ij}) \cdot Z_i^j\right)$, with

$$L_j^r(x,y) = (2\pi)^{-1/2}(x-y)^r \exp\left(-0.5\left(h_j^{-1}(x-y)\right)^2\right),$$

with h_j being the bandwidth associated with the estimation of $\hat{f}_j$

Step 2. Repeat **Step** 1 with $\hat{f}_j^0$ replaced by $\hat{f}_j$, until the convergence criterion

$$\sum_{i=1}^n \left[\hat{f}_j(Z_{ij}) - \hat{f}_j^0(Z_{ij})\right]^2 \Big/ \sum_{i=1}^n \hat{f}_j^0(Z_{ij})^2$$

is below some small threshold ε for all the $j = 1, \ldots, p$.

The algorithm above is expected to show a good behaviour if some independence assumptions are fulfilled. First, we assume that the error term in (1) is independent of the covariate vector. This implies that the model is homocedastic, and hence no extra efficiency should be gained through further weighting of the squared residuals. On the other hand, it is also assumed that the censoring variable is independent of everything else, i.e., that C gives no information on $(\mathbf{X}, Y)$. This assumption guarantees that weighting the squared, censored residuals through the jumps of the Kaplan-Meier estimator pertaining to the lifetime df, leads to consistent estimation of the $f_j(x)$'s. Besides, these independence hypotheses will be crucial for justifying the bootstrap resampling plan introduced in Section 4.

2.1 Bandwidth selection. Computational aspects

It is well known that the estimates obtained for the model heavily depend on the bandwidths $(h_1 \ldots, h_p)$ used in the local linear kernel estimates of the partial functions $(f_1 \ldots, f_p)$. The bandwidths are a trade-off between the bias and the variance of the resulting estimates. Various proposals for an optimal selection have been suggested for the additive models, yet the difficulty of asymptotic theory in a backfitting context means that nowadays optimal selection is still a challenging open problem. Cross-validation was used for the automatic choice of bandwidths.

In each of the cycles of the algorithm, the bandwidth (h_j) used to obtain the estimates $\hat{f}_j$ in equation (3) was automatically selected by minimizing the following weighted cross-validation error criterion:

$$CV_j = \sum_{i=1}^n W_i \left(Z_i^j - \hat{f}_j^{(-i)}(X_{ij})\right)^2$$

where $\hat{f}_j^{(-i)}$ is the estimate obtained without the i^{th} element of the sample.

Cross-validation implies a high computational cost, inasmuch as it is necessary to repeat the estimation operations several times in order to select the optimal bandwidths. To speed up this process, we used binning-type acceleration techniques (Fan and Marron, 1994; Wand, 1994) to obtain the binning approximations of $\hat{f}_j$ in each of the iterations of the estimation algorithm.

The binning approximations were obtained from the binning sample $\{X_r^{\bullet j}, Z_r^{\bullet j}\}$ and the weights $\{W_r^{\bullet}\}$ $(1 \leq r \leq N)$, being

$$X_1^{\bullet j} < \ldots < X_N^{\bullet j}$$

a grid of equidistant points along the j^{th} direction. Let us consider δ the distance between consecutive grids. The binning responses $Y_r^{\bullet j}$ and the binning weights $W_r^{\bullet}$ are constructed according to $W_r^{\bullet} = \sum_{i=1}^{n} W_r^{\bullet i}$ and $Y_r^{\bullet j} = \sum_{i=1}^{n} W_r^{\bullet i} Z_i^j$ with

$$W_r^{\bullet i} = W_i \left(1 - \left|X_{ij} - X_r^{\bullet j}\right| / \delta\right)_+$$

The binning approximation of the estimator $\hat{f}_j\left(x\right)$ is obtained by applying the approximations

$$s_j^r\left(x\right) \approx \sum\nolimits_{l=1}^{N} L_j^r\left(x, X_l^{\bullet j}; h_j\right) W_l^{\bullet} \text{ and } t_j^r\left(x\right) \approx \sum\nolimits_{l=1}^{N} L_j^r\left(x, X_l^{\bullet j}; h_j\right) W_l^{\bullet} Z_l^{\bullet j}$$

As in the estimation algorithm, with the binning technique the cross validation error CV_j can be approximated by:

$$CV_j \approx \sum\nolimits_{r=1}^{N} W_r^{\bullet j} \left(\hat{f}_j^{-(r)}\left(X_r^{\bullet j}\right) - \frac{Z_r^{\bullet j}}{W_r^{\bullet j}}\right)^2,$$

where $\hat{f}_{jk}^{-(r,s)}$ is obtained without the (r, s) element of the binning sample.

The finer the grid of points selected, the better the binning approximations. The choice of the number of grid points is a compromise between approximation error and computational speed. In this paper, we used $_{40}$grid points covering the range of each X_j. In practice, depending on de sample size n and on the distribution of the covariates a larger amount of grid points might be more appropriate.

3 Simulation study

A simulation study was conducted to assess the finite sample behavior of our proposed algorithm. Given the vector of covariables $\mathbf{X} = (X_1, X_2, X_3)$ in R^3, the response variable Y was generated according the model

$$\ln Y = \sum\nolimits_{j=1}^{3} f_j \left(X_j\right) + \varepsilon \tag{4}$$

with $\varepsilon \sim N\left(0,1\right)$. The censored variable C was drawn independently from a Uniform$[0, a]$. Note that the constant a determines the expected proportion of censored responses. We have chosen several values for a in order to get censoring percentages of about 0%, 15%, 33%, 50% and 67%. In all the cases, the covariates X_1, X_2 and X_3 were chosen as independent random variables distributed as Uniform$[-2, 2]$, being independent of the gaussian error and the censoring time otherwise. Then, the δ and observed variable Z were respectively given by $\delta = \mathbb{I}\left(Y \leq C\right)$ and $Z = \min(Y, C)$.One thousand independent samples $\{(\mathbf{X}_i, Z_i, \delta_i)\}_{i=1}^{n}$ of size n were generated from the model (4) under two different scenarios:

(i): $f_1\left(X_1\right) = X_1$, $f_2\left(X_2\right) = X_2$ and $f_3\left(X_3\right) = X_3$

(ii). $f_1\left(X_1\right) = X_1$, $f_2\left(X_2\right) = X_2^2$ and $f_3\left(X_3\right) = \sin\left(0.5\pi X_3\right)$

Clearly, scenario (i) is suitable for the linear model and scenario (ii) for the nonparametric additive model.

Model behaviour was evaluated for different sample sizes $n = 200$, $n = 500$ and $n = 1000$ in a new set consisting in 250 points $\{X_i^\bullet, Z_i^\bullet, \delta_i^\bullet\}_{i=1}^{250}$, generated independently from the original sample used for the estimation. Firstly, in order to compare the response predictions $\widehat{\ln Y_i^\bullet} = \sum\nolimits_{j=1}^{3} \hat{f}_j\left(X_{ij}^\bullet\right)$ we examined the mean squared error (MSE) defined as follows:

$$MSE = (1/250)\sum\nolimits_{i=1}^{250}\left(\ln Y_i^\bullet - \widehat{\ln Y_i^\bullet}\right)^2.$$

In addition, to evaluate the performance of each partial function f_j we used again the MSE criterion

$$MSE_j = (1/250)\sum\nolimits_{i=1}^{250}\left(\hat{f}_j\left(X_{ij}^\bullet\right) - f_j\left(X_{ij}^\bullet\right)\right)^2$$

Table 1 summarizes numerical average results of the considered errors over 1000 replicated samples according to each model of the corresponding scenario. The results are presented in terms of the average of the errors (along with their corresponding standard deviations) over the 1000 simulated samples.

From this table, we obtain the following conclusions. As one would expect, the linear model presents the lowest errors in scenario (i), though the additive model shows a satisfactory behavior in this situation. For scenario (ii), the shapes of f_2 and f_3 are far from linear, and therefore the $\hat{f}_2$ and $\hat{f}_3$ estimates obtained by the lineal model result in large errors. In this scenario (ii) the flexibility provided by the nonparametric estimation of the partial functions makes the additive model preferable to the parametric model.

Graphical average results are displayed in Figures 1 to 4. These figures plot the data generating functions and point-wise 95% confidence bands of the estimates $\hat{f}_1$,$\hat{f}_2$ and $\hat{f}_3$ for percentages of censored data of 0%, 50% and 80%. The good performance of the resulting estimates $\hat{f}_1$, $\hat{f}_2$, and $\hat{f}_3$ is evident for the additive model, recovering the functional forms of the corresponding true curves very successfully.

4 Application to real data

Between January, 1974, and May, 1984, the Mayo Clinic conducted a double-blinded randomized trial in Primary Biliary Cirrhosis (PBC) of the liver. A total of $n = 312$ patients agreed to participate in this clinical trial. The data were analyzed in 1986 for presentation in clinical literature (see Fleming and Harrington, 1991). Main variable of interest (the Y) was the number of days between registration and death, possibly censored because of end of study or liver transplantation. By July, 1986, 125 of the 312 patients had died, resulting in a 60% of censoring. Among 14 clinical, biochemical and histological variables, only five of them were identified as important risk factors: age at study registration, albumin (measured in gm/dl), serum bilirubin (in mg/dl), presence of edema (0=abscence, 0.5=edema present but no diuretic therapy was given, and 1=edema present despite diuretic therapy), and prothrombin time (in seconds), see Fleming and Harrington (1991), pp. 156-161. We use the notation $X_1, ..., X_5$ for these five variables in our illustration, where logarithm scale was taken for albumin, bilirubin, and prothrombin time. These data are downloadable from the R package randomSurvivalForest.

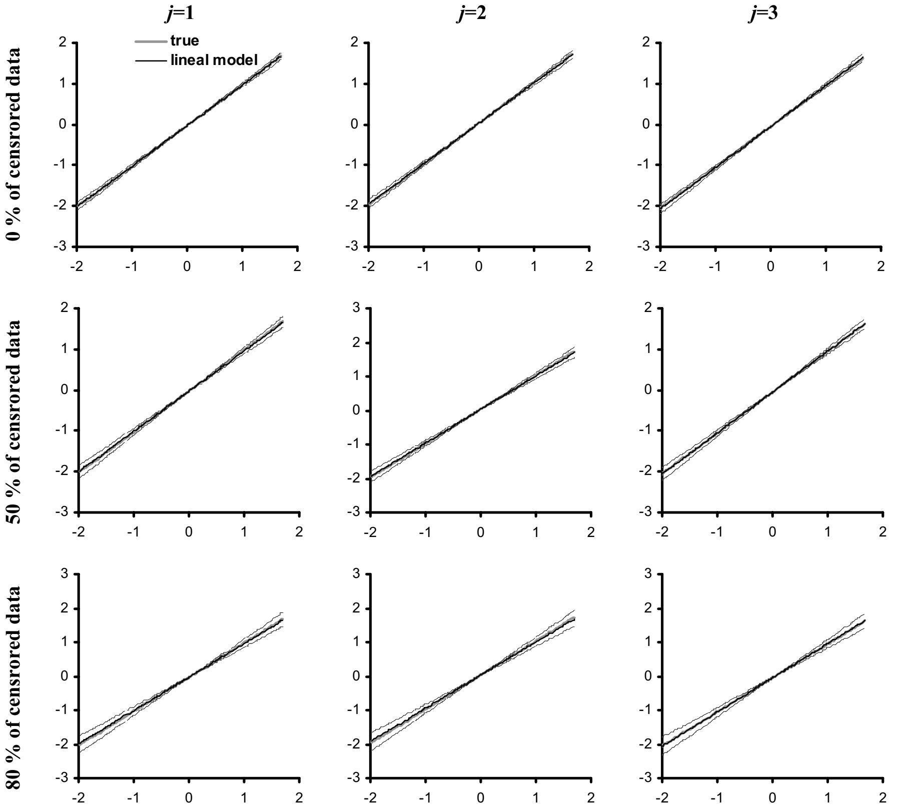

Figure 1: Data generating functions and point-wise 95% confidence bands of the estimates $\hat{f}_1$,$\hat{f}_2$ and $\hat{f}_3$. Estimates for linear model with $n = 1000$ in Scenario (i) for percentage of censored data of 0%, 50% and 80%.

We consider the additive model

$$\ln Y = \alpha + \sum_{j=1}^{5} f_j(X_j) + \varepsilon \tag{5}$$

Figure 5 depicts the linear estimates versus the nonparametric estimates for the $f_j(.)$'s, together with pointwise confidence bands based on the boostrap. We used the following bootstrap resampling plan:

Step 1. Fit Model (5), and obtain the predicted values $\ln \widehat{Y}_i = \alpha + \sum_{j=1}^{5} \widehat{f}_j(X_{ij})$ and the corresponding censored residuals $\widehat{\varepsilon}_i = \ln Z_i - \ln \widehat{Y}_i$, $i = 1, ..., n$.

Step 2. For $b = 1, ..., B$, generate the bootstrap resample $\left\{ \left(\mathbf{X}_i,\ Z_i^{*b}, \delta_i^{*b} \right) \right\}_{i=1}^{n}$

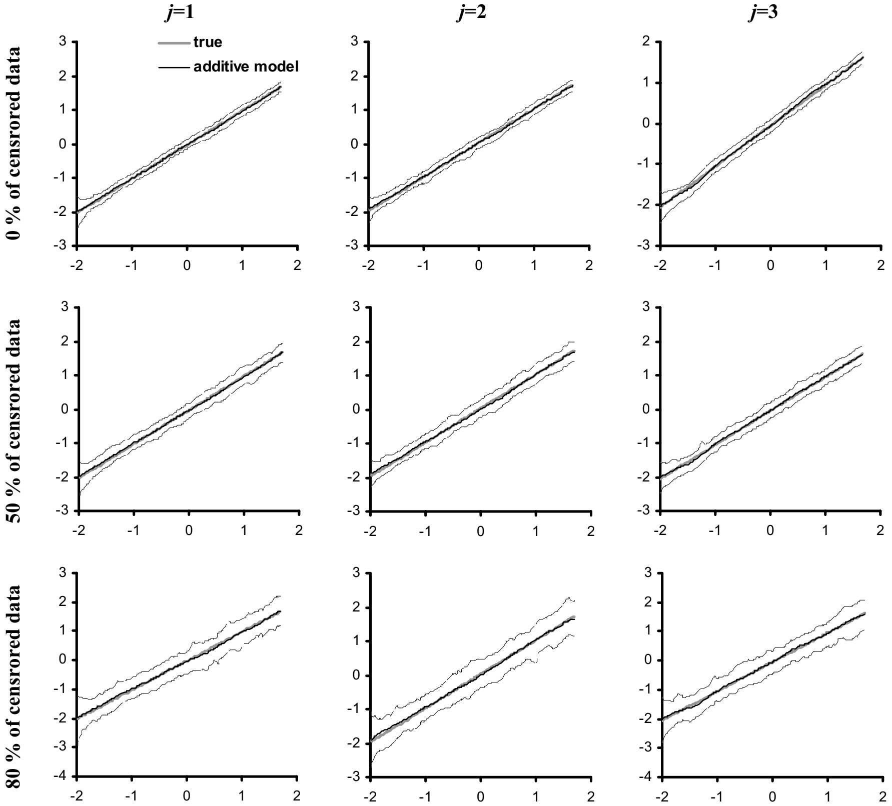

Figure 2: Data generating functions and point-wise 95% confidence bands of the estimates $\hat{f}_1$,$\hat{f}_2$ and $\hat{f}_3$. Estimates for additive model with $n = 1000$ in Scenario (i) for percentage of censored data of 0%, 50% and 80%.

where $Z_i^{*b} = \min(Y_i^{*b}, C_i^{*b})$ and $\delta_i^{*b} = \mathbb{I}(Y_i^{*b} \leq C_i^{*b})$. Here, $Y_i^{*b} = \exp\left\{\ln \widehat{Y}_i + \widehat{\varepsilon}_i^{*b}\right\}$, and the $\widehat{\varepsilon}_i^{*b}$'s and the C_i^{*b}'s are drawn from the Kaplan-Meier estimator of the residual and the censoring time df's respectively.

Note that in Step 2 the bootstrap censoring times are generated from a distribution which is not conditioned by the covariates. This is consistent with our model, under which the C and the $(\mathbf{X}, Y)$ are assumed to be independent. Similarly, since the error is independent of the covariate vector, the resampling of the residuals is performed in an unconditional way. Note that, under our model assumptions, the true residual ε is independently censored by $\ln C - \alpha - \sum_{j=1}^{5} f_j(X_j)$, and hence the Kaplan-Meier method for estimating the residual df works. Our bootstrap resampling plan is similar to that in Pardo-Fernández and van Keilegom (2006), who considered a model with

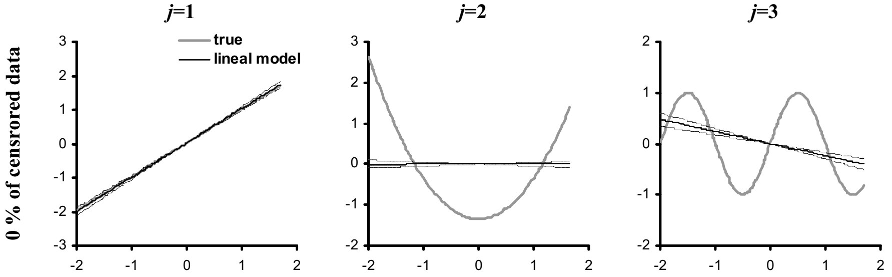

Figure 3: Data generating functions and point-wise 95% confidence bands of the estimates $\hat{f}_1$,$\hat{f}_2$ and $\hat{f}_3$. Estimates for lineal with $n = 1000$ in Scenario (ii) for noncensored data.

censoring times possibly correlated to the covariates. In our setup, however, due to our independence assumptions, we avoid the problem of choosing a bandwidth for the bootstrap. Upon completion, the $100(1-\alpha)\%$ confidence interval for $f_j(x)$ is given by

$$\left(\widehat{f}_j(x) - \widehat{f}_j^{\alpha/2}(x), \widehat{f}_j(x) - \widehat{f}_j^{1-\alpha/2}(x)\right)$$

where $\widehat{f}_j^p(x)$ represents the $1-p$ percentile of the B differences $\widehat{f}_j^{*1}(x) - \widehat{f}_j(x), ..., \widehat{f}_j^{*B}(x) - \widehat{f}_j(x)$.

Figure 5, left, shows that, under the linear additive model $f_j(x) = \beta_j x$, $j = 1, ..., 5$, the effect of albumin, bilirubin level and presence of edema is statistically significative, with the same type of correlation with survival as found through the Cox regression analysis (Fleming and Harrington, 1991). However, unlike for Cox regression, age and prothrombin time do not influence survival in a significative way. Note that our model imposes an accelerated failure time structure rather than proportional hazards, so these results should not be taken as surprising. Real findings appear when moving to the nonparametric additive regression model (Figure 5, right), under which no parametric form is assumed *a priori* for the $f_j(x)$'s. For example, the effect of bilirubin level seems to be nonlinear (also true for prothrombin time), showing a negative correlation with survival only for the largest values of the covariate (the sample median of log bilirubin is .2994), and with no correlation for the rest of values. Also interestingly, a significant positive correlation of the prothrombin time is found for the lowest quarter of covariate values (the sample first quartile is 2.303), while no significative effect is present

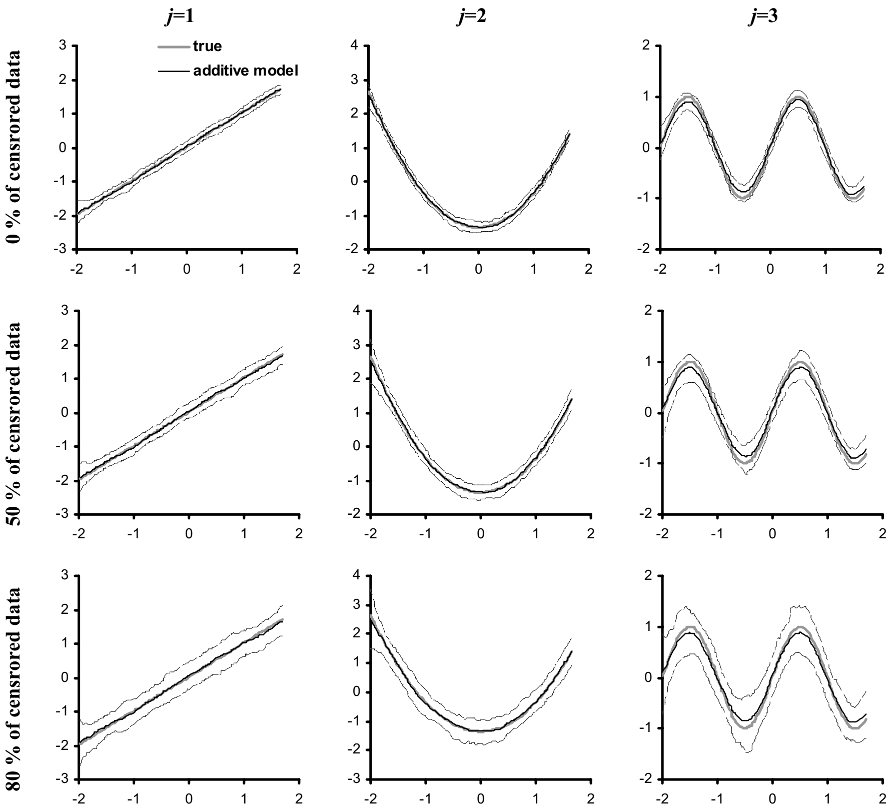

Figure 4: Data generating functions and point-wise 95% confidence bands of the estimates $\hat{f}_1$,$\hat{f}_2$ and $\hat{f}_3$. Estimates for additive model with $n = 1000$ in Scenario (ii) for percentage of censored data of 0%, 50% and 80%.

for intermediate or high levels of the covariate. This gives new insight with respect to the Cox regression analysis, for which an increasing value of the prothrombin time along its whole support would result in a higher risk of death, see the discussion in Fleming and Harrington (1991), page 161. Indeed, as recognized by Fleming and Harrington (1991), page 191 (see also the plot in their Figure 4.6.11(e) in page 192), Cox proportional hazards model could not represent the effect for prothrombin time in a proper manner. The alternative analysis offered by our additive model (which does not rely on the proportionality of the hazard functions) may have relevance to this regard.

5 Discussion

In this paper we introduce a new approach to the estimation of additive models in censored regression. Specifically, we propose an extension of the accelerated failure time regression model through additive regression, which constitutes a novelty in the context of censored regression. Weighted backfitting based on kernel smoothers has been used for estimating the model, and the smoothing windows were selected employing the cross-validation technique. Using cross validation bandwidths implies estimating the model several times, and as a consequence we have used binning acceleration techniques to speed up the estimation process.

Simulation results have shown that the proposed algorithm works well in practice, and that the additive model is a convenient alternative to linear regression in the presence of nonlinear effects. Besides, an application to real data has served for illustrating the potential advantages of our censored additive model when compared to more classical regression approaches.

Although our work is mainly focused on additive models with main effects, extensions of our methodology to factor-by-curve or curve-by-curve interactions is also possible by using bivariate kernel smoothers (Roca-Pardiñas, Cadarso-Surez and Gonzlez-Mateiga, 2005; Cadarso-Surez, Roca-Pardias and Figueiras, 2006). Moreover, the bootstrap method proposed in Section 4 could be probably used to develop statistical tests for interactions. The study of the performance of these extensions is a topic for further research.

A FORTRAN program implementing the non-parametric model estimation (with binning), and the bootstrap-based tests proposed in this paper can be obtained by contacting the first author at roca@uvigo.es.

References

Brezger A, Lang S. Generalized structured additive regression based on Bayesian P-splines. *Comput. Statist. Data Anal.* 2006; **50**: 967-991.

Cadarso-Surez, C., Roca-Pardias, J. and Figueiras, A. (2006). Effect Measures in Nonparametric Regresin with Interactions between Continuous Exposures. Statistics in Medicine 25, 603-621.

Coull B, Ruppert D, Wand M. Simple incorporation of interactions into additive models. *Biometrics* 2001; **57**, 539-545.

Eilers PHC, Marx BD. Flexible Smoothing with B-Splines and Penalties. *Statistical Science* 1996; **11**, 89-121.

Fleming, T.R. and Harrington, D.P. (1991). Counting Processes and Survival Analysis. New York: Wiley-Interscience.

Gangugli, B., Wand, M.P. (2006). Additive models for geo-referenced failure time data. Statistics in Medicine 25, 2469-2482.

Guo W. Inference in smoothing spline analysis of variance. *J. R. Statist. Soc.* B 2002; **64**, 473-491.

Hrdle, W., Huet, S., Mammen, E., Sperlich, S. Bootstrap Inference in Semiparametric Generalized Additive Models. *Econometric Theory* 2004; **20**, 265-300.

Hastie, T.J., Tibshirani RJ. Generalized Additive Models. London: Chapman and Hall, 1990

Huang, J.Z. (1999). Efficient estimation of the partly linear additive Cox model. Annals of Statistics 27, 1536-1563.

Huang, J.Z., Liu, L. (2006)Polynomial Spline Estimation and Inference of Proportional Hazards Regression Models with Flexible Relative Risk Form. Biometrics 62, 793-802.

Leurgans, S. (1987) Linear models, random censoring and synthetic data. Biometrika 74, 301-309.

Liang, H. and Zhou, Y. (1998) Asymptotic normality in a semiparametric partial linear model with right-censored data. Communications in Statistics, Theory & Methods 27, 2895-2907.

Nadaraya EA. On estimating regression. *Theory of Probability and its Applications* 1964; **10**: 186-190.

Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. J. Multivar. Anal. 73, 166-179.

Orbe, J.,Ferreira, E. and Núñez-Antón, V. (2003). Censored partial regression. Biostatistics 4, 109-121.

Pardo-Fernández, J.C. and van Keilegom, I. (2006). Comparison of regression curves with censored responses. Scandinavian Journal of Statistics 33, 409-434.

Qin, G. and Jing, B.-Y. (2000) Asymptotic properties for estimation of partial linear models with censored data. Journal of Statistical Planning and Inference 84, 95-110.

Roca-Pardias, J., Cadarso-Surez, C. and Gonzlez-Manteiga, W.(2005) Testing for Interactions in Generalized Additive Models: application to SO2 pollution data. Statistics and Computing 15, 289-299.

Ruppert D, Wand MP, Carroll RJ. Semiparametric regression. Cambridge: University Press, 2003

Sperlich S, Tjostheim D, Yang L. Nonparametric Estimation and Testing of Interaction in Additive Models. *Econometric Theory* 2002; **18**:197-251.

Stute, W. (1993) Consistent estimation under random censorship when covariables are present, Journal of Multivariate Analysis 45, 89-103.

Stute, W. (1996) Distributional convergence under random censorship when covariables are present. Scandinavian Journal of Statistics 23, 461-471.

Stute, W. (1999) Nonlinear censored regression. Statistica Sinica 9, 1089-1102.

Wand MP, Jones MC. *Kernel Smoothing.* Chapman and Hall: London, 1995.

Wahba G. Spline models for observational data. *Regl Conf. Ser. Appl. Math.* 1990; **59**.

Watson GS. Smooth Regression Analysis. *Sankhya, Series A* 1964; **26**: 359-372.

Wood SN. Thin plate regression splines . *J. R. Statist. Soc.* B 2003; **65**, 95-114.

		Scenario (i)		Scenario (ii)	
Cens.	Error	LM	AM	LM	AM
0.00	MSE	0.020	0.110	2.075	0.126
	MSE1	0.005	0.038	0.004	0.042
	MSE_2	0.005	0.037	1.479	0.041
	MSE_3	0.006	0.040	0.435	0.049
0.15	MSE	0.028	0.152	2.107	0.180
	MSE1	0.007	0.051	0.010	0.057
	MSE_2	0.006	0.047	1.496	0.062
	MSE_3	0.007	0.058	0.441	0.068
0.33	MSE	0.040	0.208	2.133	0.240
	MSE1	0.009	0.068	0.018	0.076
	MSE_2	0.007	0.062	1.508	0.083
	MSE_3	0.010	0.080	0.445	0.092
0.50	MSE	0.059	0.314	2.187	0.364
	MSE1	0.012	0.096	0.029	0.110
	MSE_2	0.010	0.097	1.532	0.133
	MSE_3	0.012	0.119	0.456	0.137
0.67	MSE	0.102	0.633	2.293	0.636
	MSE1	0.017	0.189	0.052	0.192
	MSE_2	0.017	0.242	1.572	0.216
	MSE_3	0.018	0.231	0.473	0.237
0.80	MSE	0.210	0.918	2.521	1.108
	MSE1	0.031	0.272	0.096	0.349
	MSE_2	0.036	0.256	1.675	0.392
	MSE_3	0.033	0.326	0.510	0.349

Table 1: Simulation-based averages for MSE errors of fitted LM and AM from 1000 replications for sample size $n = 200$

		Scenario (i)		Scenario (ii)	
Cens.	Error	LM	AM	LM	AM
0.00	MSE	0.010	0.057	1.690	0.071
	MSE_1	0.003	0.020	0.003	0.019
	MSE_2	0.003	0.018	1.317	0.021
	MSE_3	0.003	0.019	0.435	0.030
0.15	MSE	0.014	0.078	1.703	0.089
	MSE_1	0.004	0.027	0.004	0.025
	MSE_2	0.003	0.024	1.322	0.029
	MSE_3	0.004	0.026	0.438	0.035
0.33	MSE	0.018	0.099	1.719	0.117
	MSE_1	0.004	0.034	0.007	0.033
	MSE_2	0.004	0.030	1.331	0.038
	MSE_3	0.004	0.032	0.441	0.045
0.50	MSE	0.029	0.145	1.746	0.176
	MSE_1	0.006	0.048	0.012	0.050
	MSE_2	0.005	0.043	1.340	0.056
	MSE_3	0.006	0.048	0.447	0.063
0.67	MSE	0.049	0.249	1.778	0.262
	MSE_1	0.009	0.086	0.021	0.078
	MSE_2	0.008	0.071	1.352	0.085
	MSE_3	0.008	0.076	0.453	0.091
0.80	MSE	0.089	0.471	1.842	0.562
	MSE_1	0.013	0.154	0.031	0.192
	MSE_2	0.014	0.137	1.372	0.172
	MSE_3	0.013	0.147	0.468	0.186

Table 2: The same as in Table 1 for sample size $n = 400$

		Scenario (i)		Scenario (ii)	
Cens.	Error	LM	AM	LM	AM
0.00	MSE	0.004	0.024	1.807	0.029
	MSE_1	0.001	0.007	0.001	0.007
	MSE_2	0.001	0.008	1.412	0.008
	MSE_3	0.001	0.008	0.427	0.013
0.15	MSE	0.005	0.031	1.813	0.038
	MSE_1	0.001	0.010	0.002	0.010
	MSE_2	0.002	0.010	1.416	0.011
	MSE_3	0.001	0.010	0.429	0.017
0.33	MSE	0.008	0.042	1.819	0.049
	MSE_1	0.002	0.013	0.003	0.013
	MSE_2	0.002	0.014	1.419	0.014
	MSE_3	0.001	0.013	0.430	0.021
0.50	MSE	0.012	0.059	1.828	0.064
	MSE_1	0.002	0.018	0.004	0.017
	MSE_2	0.003	0.019	1.423	0.019
	MSE_3	0.002	0.018	0.431	0.024
0.67	MSE	0.019	0.091	1.846	0.100
	MSE_1	0.003	0.027	0.007	0.027
	MSE_2	0.004	0.027	1.431	0.031
	MSE_3	0.004	0.030	0.434	0.035
0.80	MSE	0.036	0.174	1.877	0.174
	MSE_1	0.005	0.053	0.012	0.047
	MSE_2	0.007	0.053	1.445	0.057
	MSE_3	0.006	0.053	0.440	0.058

Table 3: The same as in Table 1 for sample size $n = 1000$

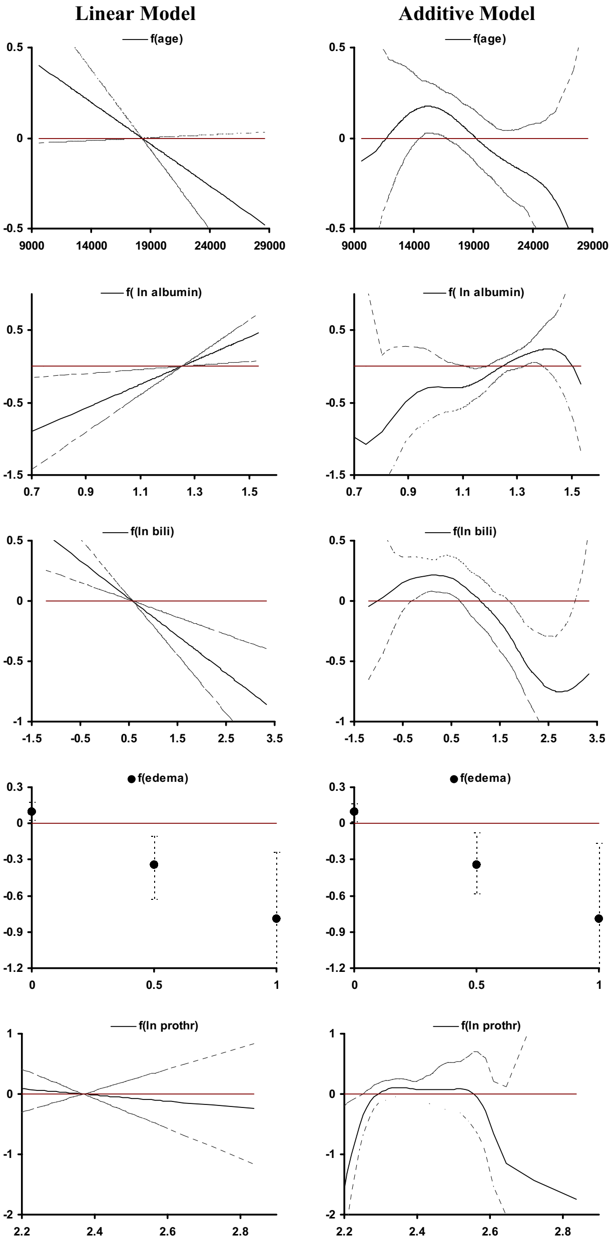

Figure 5: Estimated partial functions f_j fitted from the LM and AM together with the corresponding 95 % confidence bands.