SORT 46 (1) January-June 2022, 95-114 DOI: 10.2436/20.8080.02.119

Penalized spline smoothing using Kaplan-Meier weights in semiparametric censored regression models

Jesus Orbe[*] and Jorge Virto[*]

Abstract

In this article we consider an extension of the penalized splines approach in the context of censored semiparametric modelling using Kaplan-Meier weights to take into account the effect of censorship. We proposed an estimation method and develop statistical inferences in the model. Using various simulation studies we show that the performance of the method is quite satisfactory. A real data set is used to illustrate that the proposed method is comparable to parametric approaches when assuming a probability distribution of the response variable and/or the functional form. However, our proposal does not need these assumptions since it avoids model specification problems.

MSC: *62N02.*

Keywords: *Censored data, Kaplan-Meier weights, P-splines, semiparametric models, survival analysis.*

1. Introduction

In this paper we present a proposal for estimating regression models where the variable to be explained is censored. That is, our research context is a scenario where the values of the explanatory variables are fully known but some observations of the variable to be explained are not known because there is censored data. This problem is very common in survival or duration analyses, where the sample individuals analysed are tracked over time until the specific event studied occurs (death, failure, breakdown, etc) or the study ends. In practice, there are various types of censoring, but the most common is right censoring. There is an a large body of literature on censored data, much of which can be grouped into two main approaches: one comprising models that directly specify

[*] Department of Quantitative Methods, University of the Basque Country UPV/EHU, Bilbao, Spain
Received: October 2021
Accepted: April 2022

the effect of the explanatory variables on the variable to be explained (the most widely used of which are those known as Accelerated Failure Times (AFT) see for example Kalbfleisch and Prentice, 2002) and the other comprising hazard models, the best known and most widely applied of which is Proportional Hazard (PH), proposed by Cox (1972). In the former a regression model is specified between the logarithmic transformation of the variable to be explained and the explanatory variables. The latter specifies a relationship between the hazard function of the variable to be explained and the explanatory variables.

PH models have the advantage that the effects of the explanatory variables can be estimated without having to assume a probability distribution for the variable to be explained which is usually unknown. However they also have the drawback that the assumption of proportional hazard functions must be imposed. Another drawback of the hazard functions approach is that the effect of the explanatory variables on the variable to be explained is hard to interpret: the results obtained from Cox model fits are harder to explain to non-statisticians and provide less information than AFT-type models, which are more attractive because they can be interpreted simply and straightforwardly (Wei, 1992; Reid, 1994; Stare, Heinzl and Harrel, 2000; Swindell, 2009). Therefore, in terms of interpretability of results the linear regression model is an attractive alternative to models for hazard functions or hazard ratios. However, its main disadvantage is that the usual estimation procedure for AFT-type models requires a probability distribution to be assumed.

The proposal presented here seeks to make the modelling of this type of data more flexible without imposing restrictions or assumptions that may prove restrictive or false in practice. We also propose an approach for making inferences in this flexible model. Our proposal can be classed as an AFT type model. Several papers using this particular approach can be found in the literature which enable the regression model to be estimated with no need to choose a specific probability distribution. They consider various least squares approaches, and include the papers by Koul, Susarla and Van-Ryzin (1981) and Leurgans (1987), who propose transforming the censored variable, and those by Miller (1976), Buckley and James (1979) and Stute (1993), which present proposals with a similar approach but without transforming the variable to be explained. There is also the rank-based estimation methods approach (see for example Tsiatis, 1990; Lai and Ying, 1992; Jin et al., 2003).

These proposals represent considerable progress in the specification of the model, avoiding the biases derived from wrong choices of probability distribution. But it is possible to go even further in making these methodologies more flexible, since all these proposals consider a known parametric relationship to specify the effect of the explanatory variables on the variable to be explained. In practice, it is quite common for the functional relationship between regressor variables and outcome not to be known. One way of avoiding errors likely to lead to biased conclusions in specifying these effects is not to impose a specific parametric functional relationship between the variable to be explained and the explanatory variable, but to assume only that that relationship is a smooth function, *i.e.* to consider a nonparametric estimation of that specific effect. The estimation of nonparametric functional relationships involving non-censored data has been widely studied and various proposals have been presented in the literature. They can be grouped into two different approaches: methods based on kernel smoothers (Silverman, 1986; Härdle, 1990) and methods based on spline smoothers (Eubank, 1988; Wahba, 1990; Green and Silverman, 1994; Eilers and Marx, 1996; Wood, 2017).

Applying these nonparametric estimation techniques is not straightforward in the case of censored data, so the earlier studies must be adapted to take into account the effect of censoring in the estimation process. Our proposal falls under the spline smoothers approach in the specific context of semiparametric regression models with censored data. This semiparametric regression model has already been studied and discussed in regard to samples without censored observations. It was initially analysed by Heckman (1986) and Rice (1986) using an approach with spline smoothers and by Speckman (1988) using an approach with kernel smoothers. Several authors have investigated inference in the semiparametric regression model when the response variable is subject to right censoring. Orbe, Ferreira and Núñez Antón (2003) use an approach based on smoothing splines while Zou, Zhang and Qin (2011) and Chen et al. (2015) use penalized splines and monotone B-splines, respectively. Aydin and Yilmaz (2018) apply the ideas proposed by Koul et al. (1981) in the context of a partial linear regression model and De Uña Álvarez and Roca Pardiñas (2009) consider the use of kernel smoothers in an additive censored regression model.

A previous paper by Orbe and Virto (2018) proposes an extension of the P-splines method of Eilers and Marx (1996), which has become very popular in applications and in theoretical work and is an active area of research (Eilers, Marx and Durbán, 2015), to handle censored responses using Kaplan-Meier weights (Kaplan and Meier, 1958). But the proposal by Orbe and Virto provides no tools to allow statistical inferences to be made, and considers the case of a unique covariate. It is therefore of limited use in practice, where the response variable usually depends on a large set of explanatory variables and it is of interest to draw inferences. Here we propose an extension of that previous paper that enables the technique to be applied to more general problems where the effect of other covariates is incorporated parametrically (parametric component) in addition to the nonparametric component for modelling effects where the functional relationship is not known, that is, a semiparametric regression model. Such extension is a well-studied problem for case of uncensored data (see, for example, Heckman, 1986; Schimek, 2000; Holland, 2017). We also develop variance estimators for both the parametric and nonparametric components and provide the tools needed to develop statistical inferences in this general framework and study performance by calculating coverage probabilities of the confidence intervals for the true values of interest in several simulation studies.

The rest of the paper is organized as follows. Section 2 shows how to extend the P-splines method when the sample has censored observations and proposes a censored data version of penalized splines. Section 3 examines the methodology proposed using simulation studies. Section 4 presents an application of the method to a real data set and Section 5 concludes.

2. Methodology

The existence of censored observations is very common in survival analysis or duration analysis, where the aim is to analyse a variable that measures the duration of an event or state or the time that elapses until a specific event occurs. In other words, we consider a model that allows us to analyse the effect of certain explanatory variables on a variable to be explained T, the duration variable or usually its logarithmic transformation, where some of its observations are censored. Furthermore, we separate the effects of the explanatory variables of the model into two components: a component captures the relationship between some explanatory variables (X) and the response variable assuming a specific parametric functional form (parametric component) and the other component captures the effects of other explanatory variables (Z) whose functional form is unknown (nonparametric component) and which we leave unspecified, without assuming a particular parametric relationship. Therefore, we are considering a semiparametric regression model but in a context where the variable to be explained in the model is right-censored:

$$T_i = X_i^{\mathsf{T}}\alpha + f(Z_i) + \varepsilon_i \qquad i = 1, \ldots, n \tag{1}$$

where we assume that the values of the variable T: $t_1, \ldots, t_n$ are independent and generated with an unknown probability distribution function F. That is, we are not assuming any probability distribution for the error term. In addition some observations of that variable T are not known due to the problem known as right censoring. Therefore, what we actually observe in the sample is the variable $y_i = \min(t_i, c_i)$, where the values $c_1, \ldots, c_n$ are the values of the censoring variable C. For the censoring mechanism it will be assume: a) the lifetimes and the censoring times are independent and, b) given the lifetime, the covariates do not provide any further information as to whether censoring will take place or not, *i.e.*, $P[T \leq C|X,Z,T] = P[T \leq C|T]$ (see Stute, 1993, 1999, for a discussion of these assumptions).

We use the indicator $\delta_i = I(t_i \leq c_i)$ to show whether in particular the value t_i is observed, *i.e.*, it is not censored. In addition, X_i is the $(p \times 1)$ vector that collects the values of the p explanatory variables of the parametric component for the i-th individual, α is the $(p \times 1)$ coefficients vector of the model associated with those regressor variables, $f(Z)$ represents the nonparametric component of the model, which captures the unknown functional form of the effect of the regressor variable Z and ε is the error term satisfying $E(\varepsilon|X,Z) = 0$ and $Var(\varepsilon|X,Z) = \sigma^2$.

2.1. Estimation method

Our proposal is based on the nonparametric estimation approach proposed by Eilers and Marx (1996) together with the idea of using Kaplan-Meier weights, proposed by Stute (1993), to control the effect of censoring in the estimation of the model. Thus following this particular approach, if we want to estimate the nonparametric component of the model without assuming a particular functional form $f(\cdot)$ to the unknown effect of the regressor variable Z, we will use an approximation that rewrites or represents that effect by using a set of q B-splines type basis functions: $B_1(z),\ldots,B_q(z)$ (see, for example, Dierckx, 1993; De Boor, 2001). Thus we rewrite the unknown function as $f(z)=\sum_{j=1}^{q}\gamma_j B_j(z)$.

In order to solve the problem of choosing the number of the knots of the bases, we use the proposal of Eilers and Marx (1996) which introduces a penalty term in the estimation process of the model. This penalty term is based on the idea of previous works by O'Sullivan (1986, 1988) that propose to use a penalty term that measures the smoothness of the function through the integrated squared second derivative of the fitted function. Eilers and Marx (1996) in their proposal of the P-splines methodology suggest using, with the same idea, a different penalty term, which generalizes and simplifies the previous proposal, introducing a penalty but on the difference of the γ_j coefficients of the adjacent B-splines.

In order to account for the effect of censoring we follow the ideas of Orbe and Virto (2018) who extend the possibility of applying the P-spline methodology to the context of samples with censored observations in a simple model. Thus, to estimate the model (1) we propose to minimize the following expression:

$$\sum_{i=1}^{n} w_{[i]}\left[y_{(i)}-x_{[i]}^{\mathsf{T}}\alpha-\sum_{j=1}^{q}\gamma_j B_j(z_{[i]})\right]^2+\lambda\sum_{j=k+1}^{q}(\Delta^k\gamma_j)^2 \tag{2}$$

where $y_{(1)},\ldots,y_{(n)}$ are the ordered values of the observed variable $y_i=min(t_i,c_i)$, $x_{[i]}^{\mathsf{T}}$ is the $(1\times p)$ vector with the values of the regressors of the parametric component for the individual corresponding to the ordered observation $y_{(i)}$, $w_{[i]}$ is the Kaplan-Meier weight associated with that observation $y_{(i)}$ and this weight is calculated using the estimator $(\hat{F}_n)$ (Kaplan and Meier, 1958) of the probability distribution function F of the variable to be explained T:

$$w_{[i]}=\hat{F}_n(y_{(i)})-\hat{F}_n(y_{(i-1)})=\frac{\delta_{[i]}}{n-i+1}\prod_{j=1}^{i-1}\left[\frac{n-j}{n-j+1}\right]^{\delta_{[j]}} \tag{3}$$

without the need to assume a probability distribution for the error term, therefore a flexible methodology is used regarding to parametric assumption of the error. Furthermore $\Delta\gamma_j$ denotes the difference between the coefficients of adjacent B-splines $(\gamma_j-\gamma_{j-1})$ and $\Delta^k\gamma_j$ indicates that this difference is of order k. This difference measures the smoothness of the function $f(z)$, the larger the difference between the coefficients of adjacent B-splines the less smooth the function. Finally the parameter λ is the smoothing parameter that controls the degree of the smoothness of the estimated function in the estimation process.

The expression to minimize (2) can be rewritten in matrix form as follows:

$$(Y-X\alpha-B\gamma)^{\mathsf{T}}W(Y-X\alpha-B\gamma)+\lambda\gamma^{\mathsf{T}}D_k^{\mathsf{T}}D_k\gamma \tag{4}$$

where X is the $(n \times p)$ design matrix for the variables of the parametric component. Y is the vector of the observed variable to be explained. B is a $(n \times q)$ matrix where $B_{ij} = B_j(z_i)$. W is a $(n \times n)$ diagonal matrix with Kaplan-Meier weights. D_k is the matrix used to rewrite the Δ^k term in matrix form.

2.2. Algorithm

The optimization process of the expression (4) leads to the following equations:

$$\left(X^{\mathsf{T}}WX\right)\alpha = X^{\mathsf{T}}W(Y - B\gamma) \tag{5}$$

$$\left(B^{\mathsf{T}}WB + \lambda D_k^{\mathsf{T}}D_k\right)\gamma = B^{\mathsf{T}}W(Y - X\alpha) \tag{6}$$

In practice, the estimations of α and γ can be obtained by means of an iterative process or backfitting algorithm that iteratively solves each set of equations (5) and (6) until the convergence of the estimators is reached. We describe the algorithm process as follows:

- Step 1. In equation (6) give initial value of $\widehat{\alpha}_{(0)} = \vec{0}$ and estimate γ by $\widehat{\gamma}_{(0)} = \left[B^{\mathsf{T}}WB + \lambda D_k^{\mathsf{T}}D_k\right]^{-1}B^{\mathsf{T}}WY$.
- Step 2. Substitute γ by $\widehat{\gamma}_{(0)}$ in equation (5) and estimate α by $\widehat{\alpha}_{(1)} = [X^{\mathsf{T}}WX]^{-1}X^{\mathsf{T}}W(Y - B\widehat{\gamma}_{(0)}) = [X^{\mathsf{T}}WX]^{-1}X^{\mathsf{T}}W(I - H_c)Y$ where $H_c = B\left(B^{\mathsf{T}}WB + \lambda D_k^{\mathsf{T}}D_k\right)^{-1}B^{\mathsf{T}}W$ is the smoothing matrix for the censored case obtained from equation (6).
- Step 3. Substitute α by $\widehat{\alpha}_{(1)}$ in equation (6) and estimate γ by $\widehat{\gamma}_{(1)} = \left[B^{\mathsf{T}}WB + \lambda D_k^{\mathsf{T}}D_k\right]^{-1}B^{\mathsf{T}}W(Y - X\widehat{\alpha}_{(1)})$.
- Step 4. Iterate step 2 and step 3 until convergence is achieved.

The algorithm is considered to have converged when the difference between the GCV_c (see equation 8) of two successive iterations is less than a really small threshold: $|GCV_c(new) - GCV_c(old)| < 0.00001 \cdot GCV_c(new)$.

2.3. Choice of smoothing parameter and knots

It should be noted that in this iterative process we need to make a number of choices, such as the number of knots (K_c) and the choice of the smoothing parameter λ, in order to estimate the components of the model. The use of a penalty term in the optimization criterion makes the determination of the number of knots not a crucial decision as long as a sufficient number of knots is chosen. To choose this number of knots in samples with censored data we propose the following automatic choice criterion that takes into account the sample information available due to the existence of censored data by multiplying by one minus the proportion of censored observations:

$$K_c = round\left(\min\left(\frac{m}{4}, 40\right) \cdot (1 - PC)\right) \tag{7}$$

where m is the number of distinct values of the Z variable of the nonparametric component and PC represents the level of censoring, measured as a percentage, existing in the analysed sample. The expression (7) is a modification to the one proposed for the choice of the number of knots in Ruppert (2002) that we propose for application in contexts with censored data.

The choice of the smoothing parameter is a more relevant choice. To choose an optimal smoothing level we propose to use the following version of the generalized cross validation (GCV) criterion adapted for application in contexts with censored data:

$$GCV_c = \sum_{i=1}^{n} \frac{w_{[i]}(y_{(i)} - \hat{y}_{(i)})^2}{(n - \phi tr(H_c))^2} \tag{8}$$

where ϕ is a parameter that tries to correct for the overfitting problem that occurs when using the ordinary GCV criterion. Wood (2017) proposes to use what he refers to as the double cross validation and suggests using a value of $\phi = 1.5$. This value is justified in different ways in the literature, see for example Kim and Gu (2004) for the uncensored case and Orbe and Virto (2021) for the censored case. The performance of proposal (8) has been analysed using a simulation study and, as in the uncensored case, the choice of $\phi = 1.5$ is better in almost all situations than $\phi = 1$, with the difference increasing as the censoring increases.

2.4. Variances estimation

In this section we develop the necessary tools to perform statistical inferences for the parametric and nonparametric components.

In order to determine the variance of the parametric component, we first solve equation (6) getting $\gamma = \left(B^{\mathsf{T}}WB + \lambda D_k^{\mathsf{T}} D_k\right)^{-1} B^{\mathsf{T}}W(Y - X\alpha)$. Therefore, substituting $B\gamma = H_c(Y - X\alpha)$ in equation (5) we get $(X^{\mathsf{T}}WX)\,\alpha = X^{\mathsf{T}}W\,[Y - H_c(Y - X\alpha)]$. Solving for α we obtain $\widehat{\alpha} = [X^{\mathsf{T}}W(I - H_c)X]^{-1} X^{\mathsf{T}}W(I - H_c)Y$. Accordingly, the variance-covariance matrix of this estimator can be expressed as:

$$\begin{aligned}\widehat{Var}(\widehat{\alpha}) = \widehat{\sigma}^2 \Big\{ & \left(X^{\mathsf{T}}W(I - H_c)X\right)^{-1} X^{\mathsf{T}}W(I - H_c)(I - H_c)^t WX \\ & \left((X^{\mathsf{T}}W(I - H_c)X)^{-1}\right)^t \Big\}\end{aligned} \tag{9}$$

In a similar way, we solve equation (5) getting $\alpha = (X^{\mathsf{T}}WX)^{-1} X^{\mathsf{T}}W(Y - B\gamma)$. Plugging $X\alpha = X\,(X^{\mathsf{T}}WX)^{-1} X^{\mathsf{T}}W(Y - B\gamma) = H_p(Y - B\gamma)$, where $H_p = X\,(X^{\mathsf{T}}WX)^{-1} X^{\mathsf{T}}W$, in equation (6) we get $\left(B^{\mathsf{T}}WB + \lambda D_k^{\mathsf{T}} D_k\right)\gamma = B^{\mathsf{T}}W\,[Y - H_p(Y - B\gamma)]$. Solving for γ we get $\widehat{\gamma} = \left[B^{\mathsf{T}}W(I - H_p)B + \lambda D_k^{\mathsf{T}} D_k\right]^{-1} B^{\mathsf{T}}W(I - H_p)Y$. Accordingly, the variance-covariance matrix of this estimator can be expressed as:

$$\begin{aligned}\widehat{Var}(\widehat{\gamma}) = \widehat{\sigma}^2 \Big\{ & \left[B^{\mathsf{T}}W(I - H_p)B + \lambda D_k^{\mathsf{T}} D_k\right]^{-1} B^{\mathsf{T}}W(I - H_p)(I - H_p)^t WB \\ & \left(\left[B^{\mathsf{T}}W(I - H_p)B + \lambda D_k^{\mathsf{T}} D_k\right]^{-1}\right)^t \Big\}\end{aligned} \tag{10}$$

In order to calculate these estimated variances we need to estimate the σ^2 parameter. We propose the estimator given by the following expression:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} nw_{[i]}(y_{(i)} - \hat{y}_{(i)})^2}{n - tr(H_c) - p}$$

3. Simulation study

In this section the performance of the proposed methodology is studied using a simulation study. In order to do that we consider the next semiparametric model:

$$T_i = \alpha_1 X_{1i} + \alpha_2 X_{2i} + f(Z_i) + \varepsilon_i \tag{11}$$

where for the parametric component of the model: the variable X_1 is generated from a uniform distribution $U(0,2)$, X_2 from a uniform distribution $U(-1,3)$, being $\alpha_1 = -1$ and $\alpha_2 = 1$ the values of the coefficients. For the nonparametric component, we consider three different cases for the relationship $f(\cdot)$ between T and a relevant covariate Z, see Table 1 for the chosen functional forms and the probability distribution of the variable Z. For the distribution of the error term (ε) has been used the normal distribution $N(0,\sigma^2)$, where the value of σ^2 parameter has been chosen to obtain a similar signal/noise (SN) ratio in each example (see Table 1). In order to study the effect of censoring, we consider a censoring variable C generated independently from a uniform distribution $U(1,b)$. The value of parameter b changes to consider three different levels of censored data: 10%, 25% and 40%. Therefore, we observe $(y_1,x_{11},x_{21},z_1,\delta_1),\ldots,(y_n,x_{1n},x_{2n},z_n,\delta_n)$ a sample of size n, where $y_i = min(t_i,c_i)$ is the observed survival time, *i.e.*, the minimum between the survival time t_i and the censoring value c_i. In addition, it is known through the indicator variable $\delta_i = I(t_i \leq c_i)$ which observations are not censored. We use three sample sizes: $n = 200$, $n = 500$ and $n = 1000$. For each of the nine scenarios, three sample sizes for three levels of censorship, we consider 1000 Monte Carlo replications.

Table 1. *Three Case Studies.*

Name	z_i	$f(z_i)$	σ^2_ε	SN ratio
Case (i): Quadratic	$z_i \sim U[0,4]$	$2+4z_i-z_i^2$	0.40	3.5
Case (ii): Sinusoidal	$z_i \sim U[0,10]$	$2+exp\{sin(z_i)\}$	0.20	3.3
Case (iii): Logit	$z_i \sim U[0,1]$	$2+\dfrac{1}{1+exp\{-20(z_i-0.5)\}}$	0.06	3.3

For each of the 27 cases analysed in this simulation study we have estimated model (11) following the estimation proposal presented in the previous section, the censored P-spline estimator (CPS), where the choice of the smoothing parameter λ and the number of knots of B-splines have been chosen using formulas (8) and (7), respectively.

Tables 2, 3 and 4 present a general summary of the results obtained for each combination of censoring level and sample size in each of the three cases of functional forms studied for model (11). That is, Table 2 summarizes the estimation of case (i), where $f(z)$ is a quadratic function. The first two rows of Table 2 present the estimated Mean Square Error (MSE) of each coefficient (α_1 and α_2) of the parametric component:

$$MSE(\hat{\alpha}_p) = \frac{1}{1000}\sum_{j=1}^{1000}(\alpha_p - \hat{\alpha}_{pj})^2 \qquad p = 1,2$$

and the third row the Averaged Mean Square Error (AMSE) of the nonparametric component:

$$AMSE = \frac{1}{1000}\sum_{j=1}^{1000}\left(\frac{\sum_{i=1}^{n}(f(z_i) - \hat{f}_j(z_i))^2}{n}\right)$$

Rows four to six of Table 2 present the empirical bias and rows seven to nine the coverage probabilities of the 95% confidence intervals based on the resampling.

Tables 3 and 4 present the same information for the estimates of case (ii) and (iii), where $f(z)$ is a sinusoidal function and a logit function, respectively. Tables 2 to 4 show the good performance of the proposed method in terms of MSE and AMSE, empirical bias and coverage probabilities.

Furthermore, if we focus on the estimation of each component of model (11), we have that for case (i), quadratic function: Figure 1(a) presents the MSE estimates for the nonparametric component using different censoring levels and sample sizes, where, as can be seen, the estimates of the nonparametric component improve as the sample size increases and the level of censoring in the sample decreases. Figures 1(b) and (c) show the estimates of the coefficients of the parametric component (α_1 and α_2), where it can be seen that the coefficient estimates are good and that their accuracy also improves as the sample size increases and the level of censoring in the sample decreases. In addition, Figure 1(d) presents the mean value of the estimates of the quadratic form function compared to the true functional form to be estimated. As can be seen, the proposal we made works very well reflecting the true functional form of $f(\cdot)$. In this Figure 1(d), we can also verify the good performance of the asymptotic confidence intervals generated with the estimates of the variances proposed in the previous section. As can be seen, for a confidence level of 95%, the proposed mean confidence interval (blue lines) is consistent with the corresponding 95th percentile interval of the simulations (green lines). Finally, the coverage probabilities of the confidence intervals presented in Table 2 show that the actual coverage probability is quite close to the nominal coverage probability.

Similar results, where the good performance of our proposals can be appreciated, are obtained for case (ii), sinusoidal function, see Figures 2(a)-(d), and for case (iii), logit function, see Figures 3(a)-(d).

As suggested by the referees, we conduct additional simulations considering a normal distribution for the censoring variable and also additional simulations considering non-normal error distributions such as the Weibull distribution. The new results obtained (not shown) confirm the good performance of the proposed method and are consistent with those presented in this section.

Table 2. *Results of simulation study for the quadratic function.*

	$n = 200$			$n = 500$			$n = 1000$		
Censored %	10%	25%	40%	10%	25%	40%	10%	25%	40%
	MSE ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and AMSE ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	3.090	3.741	5.965	1.324	1.440	2.370	0.521	0.656	0.992
$\widehat{\alpha}_2$	0.722	0.906	1.581	0.275	0.302	0.541	0.121	0.181	0.259
$\widehat{f}$	9.783	12.170	21.109	4.126	5.056	8.730	2.105	2.580	4.424
	Empirical Bias								
$\widehat{\alpha}_1$	-0.00149	0.00099	0.01130	-0.00319	-0.00290	0.00042	0.00214	0.00354	0.00575
$\widehat{\alpha}_2$	0.00289	0.00206	-0.00257	0.00049	0.00039	-0.00335	0.00036	-0.00036	-0.00195
$\widehat{f}$	-0.00033	-0.00148	-0.01630	0.00239	0.00272	0.00067	-0.00232	-0.00253	-0.00575
	Coverage probabilities of the 95% confidence intervals								
$\widehat{\alpha}_1$	0.938	0.955	0.947	0.928	0.950	0.947	0.946	0.946	0.948
$\widehat{\alpha}_2$	0.945	0.946	0.934	0.941	0.960	0.943	0.960	0.926	0.957
$\widehat{f}$	0.938	0.941	0.923	0.939	0.939	0.925	0.946	0.936	0.933

Table 3. *Results of simulation study for the sinusoidal function.*

	$n = 200$			$n = 500$			$n = 1000$		
Censored %	10%	25%	40%	10%	25%	40%	10%	25%	40%
	MSE ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and AMSE ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	0.806	1.060	1.521	0.285	0.362	0.560	0.136	0.154	0.233
$\widehat{\alpha}_2$	0.189	0.266	0.376	0.062	0.087	0.132	0.035	0.044	0.064
$\widehat{f}$	4.088	5.205	7.970	1.702	2.023	3.072	0.870	1.047	1.545
	Empirical Bias								
$\widehat{\alpha}_1$	-0.00543	-0.00202	0.00083	0.00085	0.00098	0.00209	0.00060	0.00016	0.00148
$\widehat{\alpha}_2$	-0.00060	-0.00073	-0.00145	0.00051	0.00058	-0.00093	-0.00016	-0.00017	-0.00136
$\widehat{f}$	0.00674	0.00311	-0.00152	-0.00116	-0.00167	-0.00324	-0.00021	0.00010	-0.00077
	Coverage probabilities of the 95% confidence intervals								
$\widehat{\alpha}_1$	0.944	0.936	0.925	0.956	0.955	0.944	0.948	0.956	0.940
$\widehat{\alpha}_2$	0.949	0.930	0.938	0.952	0.938	0.944	0.944	0.943	0.962
$\widehat{f}$	0.930	0.927	0.918	0.932	0.941	0.932	0.942	0.938	0.941

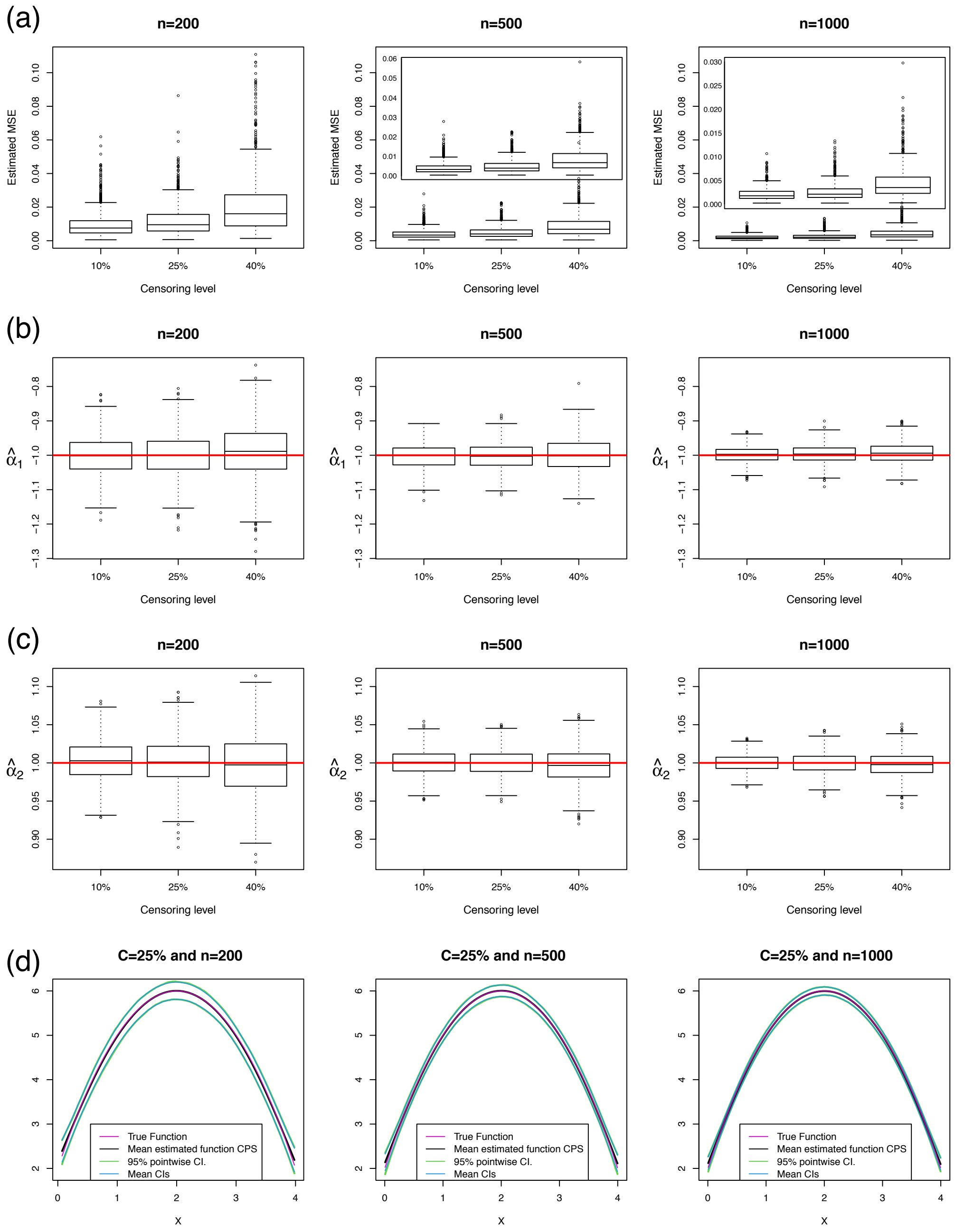

Figure 1. *Results of simulation study for the quadratic function. (a) Mean square errors for the nonparametric part using different censoring levels and sample sizes. (b) $\hat{\alpha}_1$. (c) $\hat{\alpha}_2$. (d) Mean value of the estimates of the quadratic form function compared to the true functional form to be estimated.*

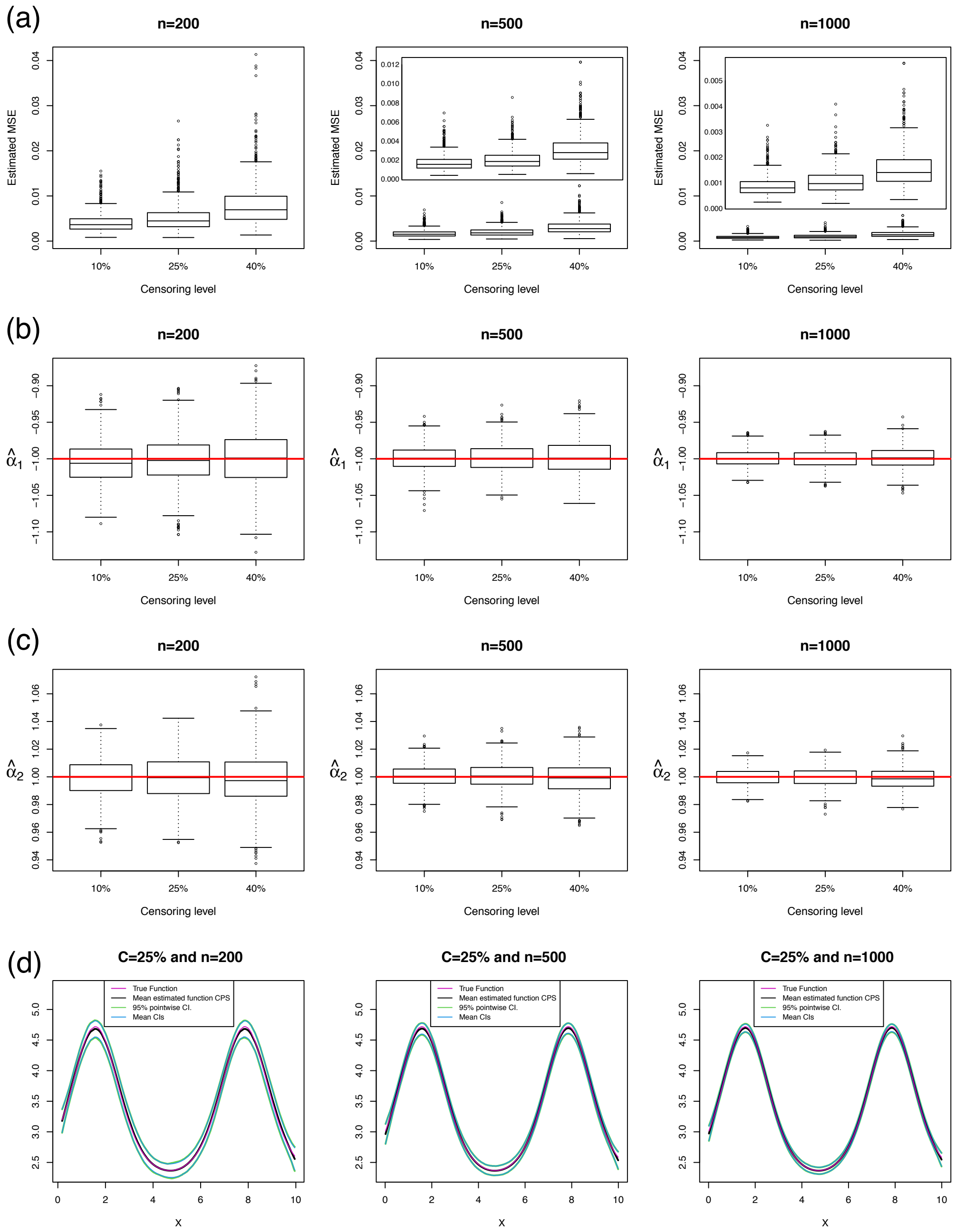

Figure 2. *Results of simulation study for the sinusoidal function. (a) Mean square errors for the nonparametric part using different censoring levels and sample sizes. (b) $\hat{\alpha}_1$. (c) $\hat{\alpha}_2$. (d) Mean value of the estimates of the sinusoidal form function compared to the true functional form to be estimated.*

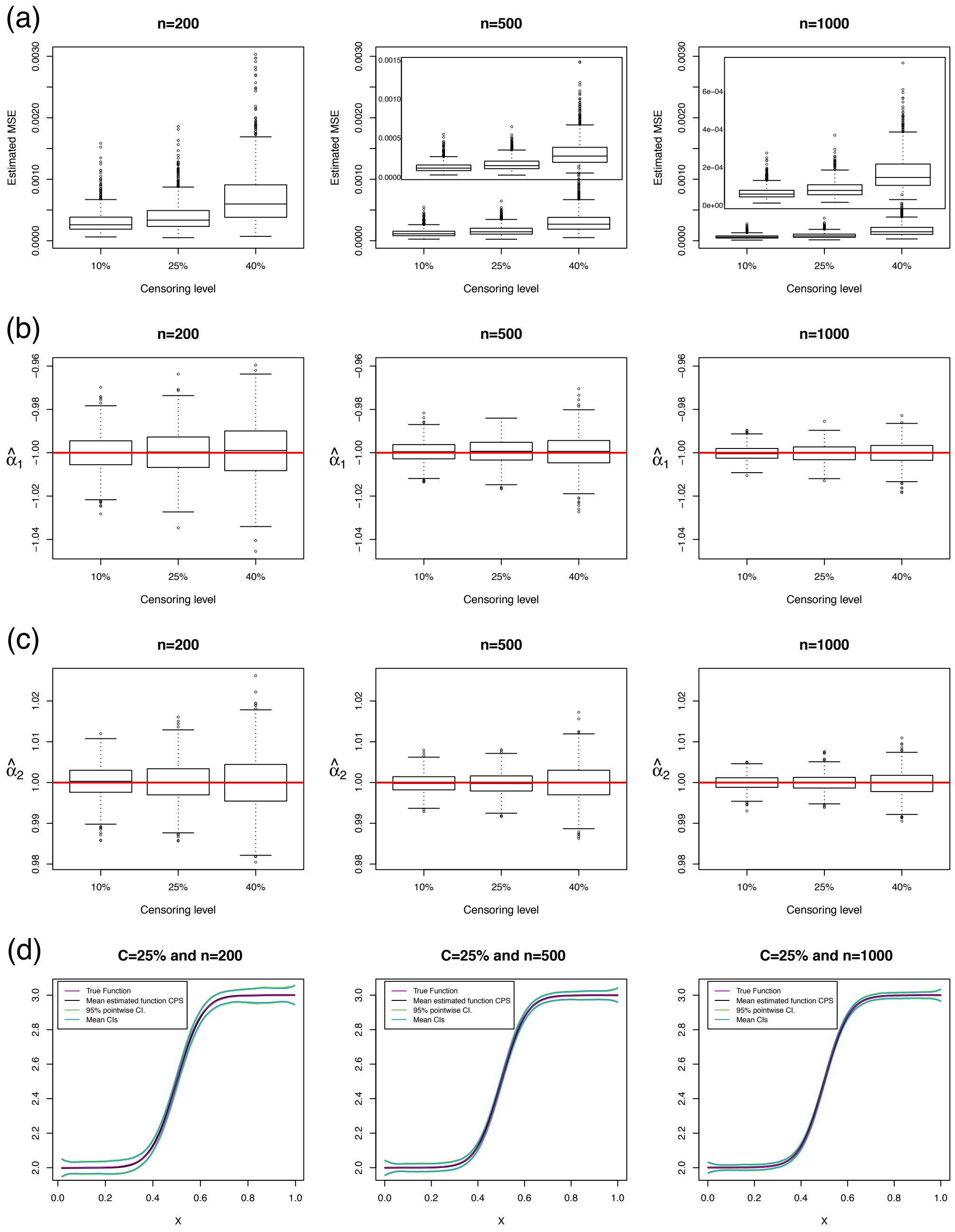

Figure 3. *Results of simulation study for the logit function. (a) Mean square errors for the nonparametric part using different censoring levels and sample sizes. (b) $\hat{\alpha}_1$. (c) $\hat{\alpha}_2$. (d) Mean value of the estimates of the logit form function compared to the true functional form to be estimated.*

Table 4. *Results of simulation study for the logit function.*

	$n = 200$			$n = 500$			$n = 1000$		
Censored %	10%	25%	40%	10%	25%	40%	10%	25%	40%
	MSE ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and AMSE ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	0.072	0.098	0.172	0.024	0.036	0.064	0.011	0.016	0.027
$\widehat{\alpha}_2$	0.017	0.025	0.046	0.006	0.008	0.019	0.003	0.004	0.009
$\widehat{f}$	0.309	0.397	0.710	0.128	0.164	0.311	0.065	0.085	0.169
	Empirical Bias								
$\widehat{\alpha}_1$	0.00012	0.00029	0.00101	0.00042	0.00061	0.00035	-0.00029	-0.00022	-0.00011
$\widehat{\alpha}_2$	0.00026	0.00015	-0.00008	-0.00015	-0.00026	-0.00002	-0.00002	-0.00003	-0.00014
$\widehat{f}$	-0.00037	-0.00038	-0.00098	-0.00022	-0.00029	-0.00040	0.00035	0.00020	0.00014
	Coverage probabilities of the 95% confidence intervals								
$\widehat{\alpha}_1$	0.944	0.955	0.933	0.953	0.944	0.941	0.946	0.941	0.948
$\widehat{\alpha}_2$	0.950	0.930	0.924	0.939	0.947	0.938	0.957	0.938	0.956
$\widehat{f}$	0.944	0.939	0.916	0.943	0.938	0.93 0	0.949	0.941	0.938

4. Empirical application: PBC data

The Mayo Clinic Primary Biliary Cirrhosis dataset contains information from 418 Mayo Clinic patients with primary biliary cholangitis (PBC), previously called primary biliary cirrhosis, an autoimmune disease of the liver. The first 312 cases in the dataset participated in a Mayo Clinic trial in PBC conducted between 1974 and 1984 comparing the drug D-penicillamine (treatment) with a placebo. The dataset provides information about the observed survival time from the date of registration in the trial and a large number of clinical, biochemical, serologic and histologic variables such as patient's age at first diagnosis, severity of edema (0 no edema, 0.5 moderate and 1 for severe edema), blood values related to liver function such as bilirubin, albumin, alkaline phosphotase and prothrombin time amid other explanatory variables, and an indicator of patient status (dead or alive) in July 1986. The dataset can be downloaded from the R package *survival* (Therneau, 2021; R Core Team, 2018). The additional cases are from an independent set of 106 Mayo Clinic primary biliary cholangitis patients who were elegible for the trial but declined to participate. This dataset has been previously used, for example, in Dickson et al. (1989), Therneau and Grambsch (2000) and Fleming and Harrington (2005), in censored regression models.

The studies by Therneau and Grambsch (2000) and Fleming and Harrington (2005) deal with the relationship between the covariates and the survival response variable. They conclude that age, edema score, bilirubin and albumin logarithms and prothrombin time are the variables that best explain patient survival. In addition, these studies analyse the need for transformations of the continuous variables in the proposed model

Table 5. *Estimate and standard deviation (SD) of estimated parameters for the Mayo Clinic Primary Biliary Cirrhosis dataset from AFT, Stute and CPS methods.*

	age	edema	trt	log(albumin)	log(bili)
AFT	-0.0246	-0.7692	-0.0627	1.4880	-0.5356
	(0.0065)	(0.2303)	(0.1273)	(0.5268)	(0.0694)
Stute	-0.0166	-0.9249	-0.0950	1.6161	-0.3028
	(0.0076)	(0.3489)	(0.1371)	(0.6015)	(0.0732)
CPS	-0.0168	-0.9163	-0.0991	1.6197	-0.3061
	(0.0064)	(0.1900)	(0.1291)	(0.4578)	(0.0633)

concluding that the relationship between prothrombin time (protime) and patient survival is likely to be non-linear.

In this application we incorporate the protime variable into the model in a flexible way only assuming that prothrombin time enters in the model via some unknown smooth function $f(\cdot)$:

$$log(T) = \alpha_1 + \alpha_2 age + \alpha_3 edema + \alpha_4 trt + \alpha_5 log(albumin) + \alpha_6 log(bili) + f(protime) + \varepsilon \tag{12}$$

We estimated model (12) using the censored P-spline method proposed in section 2. To evaluate the performance of the censored P-spline estimator, a quadratic relationship between the logarithm of survival and the protime variable has been proposed as an alternative, *i.e.*, $f(protime) = \alpha_7 protime + \alpha_8 protime^2$ in equation (12). Assuming that this parametric specification is correct, two methodologies known and proposed in the literature on survival analysis can be used to fit the model (12). These estimators can be used as a benchmark to evaluate the performance of the censored P-spline method proposed. The first and more restrictive approach is the parametric Accelerated Failure Time (AFT) methodology (Kalbfleisch and Prentice, 2002), based on the restricted assumptions of knowing the probability distribution of the response variable and the functional form relating the protime variable and patient survival, that estimates the α coefficients of the model using the maximum likelihood estimator. Thus, considering an AFT lognormal model, we estimate the α coefficients assuming a normal probability distribution. The second methodology, proposed by Stute (1993), is less restrictive in that it does not need the assumption of the probability distribution of the response variable, but it also trusts the quadratic functional form. That is, it needs to know the form of the relationship between the response variable and the covariate. This methodology estimates coefficients using weighted least squares via Kaplan-Meier weights (Stute, 1993).

Table 5 presents the estimates of the parametric components of the model (12) using these three methods. It can be seen that all three methods generate similar estimates and result in a biologically reasonable model estimate. As previously reported in the literature, all three methods agree that treatment with the drug D-penicillamine (treatment) has no significant effect on patient survival.

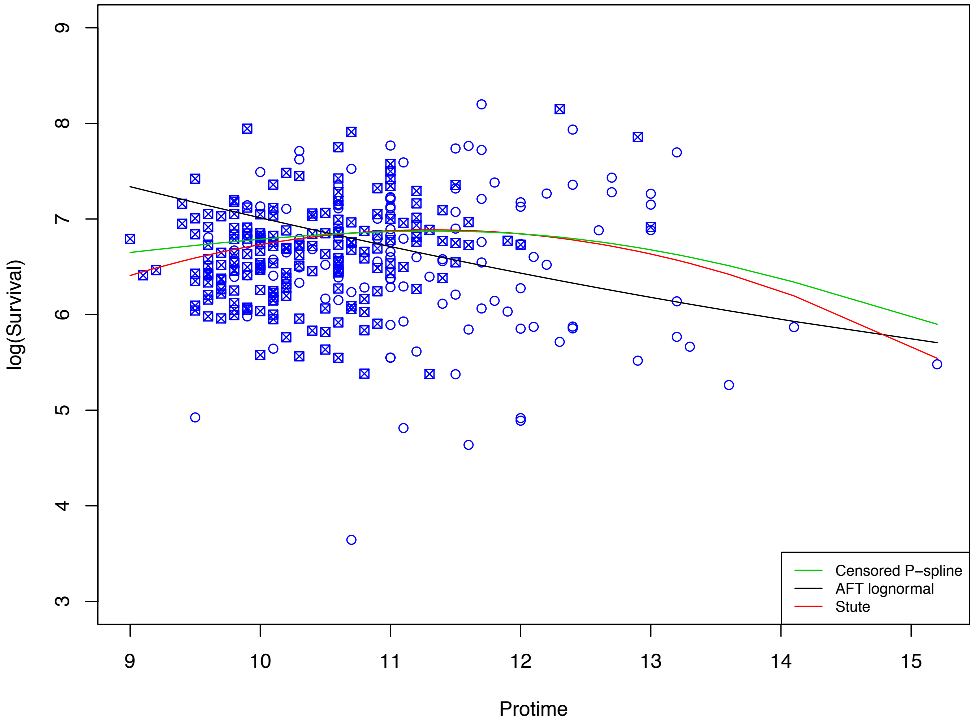

Figure 4. *Estimated relationship using three methodologies: AFT lognormal, Stute's approach and CPS estimator*

Figure 4 shows the estimates of the unknown function $f(protime)$ for the three approaches with the scatterplot of observed log survival time versus prothrombin time. Patients indicated by ○ are dead and those indicated by ☒ are alive in July 1986; that is, the dead patients have uncensored survival times and the live patients have censored survival times.

In conclusion, the AFT methodology and Stute‘s proposals performance depends on the correct specification of the relationship between the duration and the protime variable. In this application it seems that the relationship between log survival and prothrombin time is quadratic, so both these methodologies perform reasonably well. Our proposal does not need to assume a specific parametric functional form and, however, it adequately estimates the relationship obtaining very similar results to the previous ones. However, if the functional form had been wrongly chosen these parametric methods would have led to a serious problem of incorrect specification and therefore to wrong conclusions. Therefore, we can see our approach as a robust solution to misspecification of the model.

5. Discussion and conclusion

In this paper, we have proposed an estimation method in the context of censored semiparametric models based on the P-spline approach of Eilers and Marx (1996) using Kaplan-Meier weights to take into account the effect of censorship. We present an extension of the estimation methodology proposed by Orbe and Virto (2018) to a context with more than one explanatory variable, which is very useful from a practical point of view. Furthermore, we develop the necessary tools to perform statistical inferences in this general framework, providing, for example, confidence intervals for both the nonparametric component and the coefficients associated with the regressors of the parametric component. The simulation studies conducted illustrate the good performance of the estimation method which satisfactorily estimates both the nonparametric component and the coefficients associated with the parametric part in the various examples studied. Furthermore, the accuracy of estimates improves as the censored level reduces and the sample size is increased. The coverage probabilities of the confidence intervals proposed have been calculated in several simulation studies and it has been found that the actual coverage probability is quite close to the nominal coverage probability in all the scenarios analysed.

The application to real data serves to illustrate the potential advantages of our proposal which is comparable with the parametric method AFT and Stute's approach when the functional form chosen is correct. Otherwise, it must be mentioned that if the functional form or the probability distribution are wrongly chosen this would lead to a serious problem of incorrect specification of the model and therefore to incorrect conclusions. The proposed method would be more flexible and robust as it does not need to impose a specific probability distribution for the response variable, nor assume a functional form for the relationship between the censored response variable and the covariate, which are usually unknown in practice. Therefore, its application in samples with censored data is particularly useful in contexts of survival or duration analysis where censored observations are common.

Funding

This study was supported by the Basque Government and the Spanish Research Agency of the Ministry of Science and Innovation under research grants UPV/EHU Econometrics Research Group IT1359-19 and PID2019-105183GB-I00.

References

Aydin, D. and Yilmaz, E. (2018). Modified estimators in semiparametric regression models with right-censored data. *Journal of Statistical Computation and Simulation*, 88:1470–1498.

Buckley, J. J. and James, I. R. (1979). Linear regression with censored data. *Biometrika*, 66:429–436.

Chen, W., Li, X., Wang, D., and Shi, G. (2015). Parameter estimation of partial linear model under monotonicity constraints with censored data. *Journal of the Korean Statistical Society*, 44:410–418.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34:187–202.

De Boor, C. (2001). *A Practical Guide to Splines, revised version*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York.

De Uña Álvarez, J. and Roca Pardiñas, J. (2009). Additive models in censored regression. *Computational Statistics and Data Analysis*, 53:3490–3501.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, 10:1–7.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, 11:89–121.

Eilers, P. H., Marx, B. D., and Durbán, M. (2015). Twenty years of p-splines. *SORT-Statistics and Operations Research Transactions*, 39(2):149–186.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.

Fleming, T. R. and Harrington, D. P. (2005). *Counting Processes and Survival Analysis*. John Wiley & Sons, Hoboken: New Jersey.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, volume 58 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

Härdle, W. (1990). *Applied Nonparametric Regression*, volume 19 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48:244–248.

Holland, A. D. (2017). Penalized spline estimation in the partially linear model. *Journal of Multivariate Analysis*, 153:211–235.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, 90:341–353.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kim, Y. J. and Gu, C. (2004). Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:337–356.

Koul, H., Susarla, V., and Van-Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, 9:1276 – 1288.

Lai, T. L. and Ying, Z. (1992). Linear rank statistics in regression analysis with censored or truncated data. *Journal of Multivariate Analysis*, 40:13–45.

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

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, 63:449–464.

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

Orbe, J. and Virto, J. (2018). Penalized spline smoothing using Kaplan-Meier weights with censored data. *Biometrical Journal*, 60:947–961.

Orbe, J. and Virto, J. (2021). Selecting the smoothing parameter and knots for an extension of penalized splines to censored data. *Journal of Statistical Computation and Simulation*, 91:1–33.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, 1:502–527.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, 9:363–379.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reid, N. (1994). A conversation with sir david cox. *Statistical Science*, 9:439–455.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics and Probability Letter*, 4:203–208.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11:735–757.

Schimek, M. G. (2000). Estimation and inference in partially linear models with smoothing splines. *Journal of Statistical Planning and Inference*, 91:525–540.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50:413–436.

Stare, J., Heinzl, H., and Harrel, F. (2000). On the use of buckley and james least squares regression for survival data. In Ferligoj, A. and Mrvar, A., editors, *New Approaches in Applied Statistics*, volume 16, pages 125–134. Metodološki zvezki, Ljubljana: Eslovenia.

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

Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica*, 9:1089–1102.

Swindell, W. R. (2009). Accelerated failure time models provide a useful statistical framework for aging research. *Experimental Gerontology*, 44:190–200.

Therneau, T. M. (2021). *A Package for Survival Analysis in R*. R package version 3.2-11.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, 18:354–372.

Wahba, G. (1990). *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11:1871–1879.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Texts in Statistical Science Series. CRC press, Boca Raton: Florida.

Zou, Y., Zhang, J., and Qin, G. (2011). A semiparametric accelerated failure time partial linear model and its application to breast cancer. *Computational Statistics and Data Analysis*, 55:1479–1487.