Survival Analysis Using a Censored Semiparametric Regression Model

Jesus Orbe, Eva Ferreira and Vicente Núñez-Antón [1]

Abstract. In this work we study the effect of several covariates X on a censored response variable T with unknown probability distribution. A semiparametric model is proposed to consider situations where the functional form of the effect of one or more covariates is unknown. We provide its estimation procedure and, in addition, a bootstrap technique to make inference on the parameters. An application with a real dataset is presented, as well as some simulation results, to demonstrate the good behavior of the proposed estimation process and to analyze the effect of the censorship. This new model has an important application field in reliability, survival or lifetime data analysis.

Keywords: Duration Models, Censorship, Kaplan-Meier, Bootstrap, Nonparametric Estimation.

[1]The authors would like to thank Winfried Stute for valuable comments and discussion. This work was partially supported by Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU) and Dirección General de Enseñanza Superior del Ministerio Español de Educación y Cultura under research grants UPV 038.321-HA129/99 and PB98-0149.
Correspondence address: Dpto. de Econometría y Estadística. Facultad de Ciencias Económicas y Empresariales. Universidad del País Vasco-Euskal Herriko Unibertsitatea. Avenida Lehendakari Agirre 83; 48015 Bilbao, Spain. Phone: (34)946013842, Fax: (34)946013754, E-mail: etporlij@bs.ehu.es

1. Introduction: Traditional Methodologies

In survival, duration or reliability studies, it is of interest to analyze the length of time spent until some particular event happens (e. g. death or failure). This type of studies is very common in fields such as Medicine, Engineering or Economics. The analysis of duration data involves working with data with some special characteristics:

(i) Censorship, since at the end of the study the complete duration of some of the observations is unknown.

(ii) Asymmetric distributions, usually presenting a positive asymmetry, which implies that the assumption of a normal distribution is not adequate. Thus, we have to consider other more appropriate distributions such as, for example, the Weibull, exponential or Gamma distributions.

As a result, the traditional methods applied in standard problems in Statistics cannot be used. In order to solve this issue and taking into account the special characteristics of this kind of data, several specific methodologies, suitable for these data, have been developed.

Let T be a random variable measuring the time until some event happens, that is, the duration variable, and let X represent the relevant covariates considered to explain T. There are two big classes of regression models that analyze the dependence between X and T. The proportional hazards models (Cox, 1972) and the accelerated failure time models (see, e.g., Lawless, 1982).

In the Cox model, we have the following specification for the hazard function

$$\lambda(t, x) = \lambda_0(t)h(x, \beta),$$

where $h(x, \beta)$ is usually considered as $\exp(x\beta)$ and $\lambda_0(t)$ is known as the baseline hazard

function. Thus, the effect of the covariates in this model is multiplicative on the baseline hazard. The advantage of this model, and the main reason for its extensive use, is the possibility to estimate the parameters of interest without any assumption on the distribution of the duration variable. That is, there are no parametric restrictions on the functional form of the baseline hazard function. However, the assumption of a proportional hazard function for the different individuals is very restrictive, and, in some cases, this proportionality is not verified by the data. Therefore, for these cases, this model should not be used. The estimation of this model can be carried out using the partial likelihood function (Cox, 1975).

The other important class of models is the accelerated failure time models. In these models, the hazard function is modelled as

$$\lambda(t, x) = \lambda_0(t \cdot h(x, \beta))h(x, \beta).$$

Here, we have the multiplicative effect on the baseline hazard and a direct effect on the duration accelerating or decelerating the pass to another stage (e.g., failure or death). In addition, if we take $h(x, \beta) = \exp(-x\beta)$, we can rewrite the model considering a direct relation between the duration and the covariates. That is,

$$\log(T) = x\beta + \epsilon.$$

Usually, the estimation of this model is carried out assuming a distribution for the duration and maximizing the corresponding likelihood function, where the contribution of a censored observation is given by the survival function and the contribution of an uncensored one given by the density function.

Only two parametric models, the Weibull regression model and, as a particular case, the exponential regression model, can be considered within these two classes of traditional duration regression models.

The rest of the paper is outlined as follows. In Section 2, we present a flexible alternative for the traditional methodology. An illustration of this method is given in Section 3. In Section 4, we propose a new model, a censored partial regression model, more flexible than the one presented in Section 2 and provide the details for its estimation procedure. Section 5 contains a new proposal to make inference in the censored partial regression model. In Section 6, the new methodology is illustrated with an application to a real dataset. Section 7 provides some simulation results to demonstrate the good behavior of the proposed estimation process and Section 8 presents some discussion about the methods proposed here.

2. A Flexible Alternative Methodology

Stute (1993) presents a new methodology for regression with censored data which requires very general hypotheses and where the estimators can be obtained using weighted least squares.

We now briefly describe this methodology. Let us assume that $T_1, \ldots, T_n$ are independent observations from some unknown distribution function F and, because of the censoring, not all of the T's are available. That is, rather than observing T_i, we observe

$$Y_i = \min(T_i, C_i), \qquad \delta_i = \begin{cases} 1; & \text{if } T_i \leq C_i \\ 0; & \text{if } T_i > C_i \end{cases},$$

where $C_1, \ldots, C_n$ are the values for the censoring variable C, which is independent of the duration variable T, and δ_i is the indicator for the censoring variable. In addition, X_i represents the k-dimensional vector of covariates for the i-th individual. The relation between the covariates and the duration is then given by

$$T_i = X_i\beta + \epsilon_i \quad \text{with} \quad \mathrm{E}[\epsilon_i | X_i] = 0. \tag{1}$$

The estimator of β can be obtained by minimizing

$$\sum_{i=1}^{n} W_{in}[Y_{(i)} - X_i\beta]^2,$$

where $Y_{(i)}$ is the i-th ordered value of the observed response variable Y and W_{in} are the Kaplan-Meier weights. These weights can be calculated using the expression

$$W_{in} = \hat{F}_n(Y_{(i)}) - \hat{F}_n(Y_{(i-1)}) = \frac{\delta_i}{n-i+1} \prod_{j=1}^{i-1} \left[\frac{n-j}{n-j+1}\right]^{\delta_j}, \tag{2}$$

where $\hat{F}_n$ is a Kaplan-Meier estimator (Kaplan and Meier, 1958) of the distribution function F. These weights can also be calculated using the redistribute to the right algorithm presented by Efron (1967). This algorithm can be described in the following steps: (i) put in order the observed duration variable; (ii) give the same weight to all observations; (iii) take the smallest observation and, if it is a censored observation, assign a zero weight to it and distribute its weight among the rest of larger observations. However, if the observation is not censored, it keeps the weight it had been assigned. Finally, step (iv) indicates to repeat this process to all observations starting with the smallest one and ending with the largest one.

In this way, the estimator for β is given by

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y,$$

where $Y = (Y_{(1)}, \ldots, Y_{(n)})^T$, W is a diagonal matrix with the Kaplan-Meier weights on its main diagonal and $X = [X_1^T, X_2^T, \ldots, X_n^T]^T$ is the design matrix or matrix of covariates. Stute (1993) studies the consistency of this estimator, and Stute (1996a) its asymptotic normal distribution.

Model (1) can be considered within the class of accelerated failure time models. However, it allows the estimation without assuming any distribution for the duration and, in

addition, it does not require the assumption of proportional hazard functions. Therefore, this model is an interesting alternative to the previous ones.

3. Application: A First Approach to Study the Survival Time for AIDS Patients

As an illustration of the methodology described above, we use a dataset that contains information giving the survival time for AIDS diagnosed patients who lived in the Basque Autonomous Community and the Autonomous Community of Navarra in Spain. We have a sample of 461 patients diagnosed with AIDS from 1984 until December 31, 1990. The duration variable under study measures the survival time for the patient from the illness diagnosis time up to death or up to the end of the study (censored observations). Unlike most of the research done in this area, which is centered on the study of the duration for the incubation period, we are interested in the study of the duration of the last stage of the illness. The evolution of the HIV virus has three stages. The first one is known as the "pre-antibody" stage and it is the shortest one with a duration of only several months (approximately 50% of the patients generate antibodies two months after the infection date). This stage goes from the infection date to the development of antibodies or seroconversion point. During this period the patient is classified as seronegative. The second stage, the "incubation" stage, is the largest one (approximately half of the infected people develop the illness before 10 years). This period starts with the seroconversion and goes until the diagnosis of AIDS. Along this period, the individual is classified as seropositive. Finally, the third stage gives the survival time from the AIDS diagnosis time up to death or up to the end of the study. This period starts when the individual develops some of the illnesses classified into the ones related to AIDS.

As of December 31, 1992, there were 447 patients that had died and 14 patients that

had survived. Therefore, we have 14 censored observations.

In order to describe the survival time, we have several covariates which contain the characteristics of the individuals: sex (**Sex**), age at diagnosis (**Age**), transmission category, disease at AIDS diagnosis and the period of diagnosis (**Period**). The transmission category is coded using five dummy variables: **T-Sex, T-Drug, T-Blood, T-Moth-child** and **T-Others**, taking value one if the transmission via is the indicated, and zero otherwise. The disease at AIDS diagnosis is coded using three dummy variables: **Disease1**, takes value one if the patient has been diagnosed with AIDS through an opportunistic infection; **Disease2**, takes value one if the AIDS diagnosis is produced by a Kaposi's sarkoma or some lymphoma; and **Disease3**, takes value one if the patient has been diagnosed through an HIV encephalopathy or a HIV wasting syndrome. Finally, the period of diagnosis (**Period**) takes value one if the diagnosis took place after 1987. The last variable is used to capture the effect of the introduction of the AZT treatment (that started its administration in the middle of 1987) on the survival of the patients.

We want to study the effect of these explanatory variables on the survival time from the moment of diagnosis, without assuming any distribution for the duration. In order to do this, we have decided to use the flexible model in (1) with the logarithmic transformation of the duration. Table 1 summarizes the most relevant results obtained for our dataset. The standard deviations of the estimated coefficient (SDEV) are calculated using a jackknife estimator. Stute (1996b) shows the consistency of this jackknife estimator for the variance.

Note that the age of the patient has a significant negative effect on the survival time of the patient. That is, the older the patient is at the moment of diagnosis, the shorter the length of his survival time is. Other relevant variable is the period of diagnosis. We find out that patients whose diagnosis is posterior to 1987 have larger survival times. Thus, this can be an indicator of a beneficial effect of the AZT treatment to lengthen the survival

time of the patients. The rest of the covariates are not significant to explain the survival time of the patient. In addition we point out that the application of this methodology gives us the same results obtained in a previous study (Orbe et al., 1996), where we used the traditional methodologies described in Section 1. These same conclusions have been obtained by other authors as reflected in a synthesis of results presented by Brookmeyer and Gail (1993).

Table 1: Estimates of β for the parametric model

VARIABLE	COEF	SDEV
Constant	1.2574	0.4306
Sex	0.0579	0.1360
Period	0.2435	0.1226
Disease1	-0.0774	0.2266
Disease2	-0.2002	0.2995
T-Sex	-0.0554	0.2701
T-Drug	0.0106	0.2337
T-Blood	0.0816	0.3015
T-Moth-child	0.6459	0.5085
Age	-0.0160	0.0069

In this model, we have tried to capture the effect of the introduction of the AZT treatment using a dummy variable, which divides the period of study into two parts, after and before 1987. We consider that this specification is quite restrictive and not very flexible. It seems more logical to assume that this effect would be more gradual than the one specified using a dummy variable. As a result of this, we propose to model this effect nonparametrically; that is, without specifying any functional form for the relationship between the period of diagnosis variable and the duration variable. In addition, this new proposal allows us to study the total evolution of the diagnosis period effect on the duration variable. The need to have a more flexible specification leads us to propose a new model,

the censored partial regression model, which will be described in the next section.

4. The Censored Partial Regression Model

This model generalizes model (1). Thus, our proposal extends considerably the application field of the previous model without assuming any probability distribution for the duration, without requiring proportionality of the hazard functions and modelling the direct effect of the covariates on the duration. It allows us to model situations where we do not know the functional form of the effect of one covariate on the response variable, or situations where the assumption of a lineal dependence, or any other different one between some covariate and the duration variable, is a restrictive assumption, or, even, it does not make any sense. The proposed model is a semiparametric one; that is, it is a model where the effect of the covariates can be separated in two components: a parametric one, as in model (1), and a nonparametric one, where we do not specify a specific functional form for the effect of the covariate on the duration. Taking this into account, we introduce a smooth function $h(\cdot)$ to model the effect of some covariate R on the duration. Thus, the censored partial regression model proposed can be written as

$$\ln T_i = X_i\beta + h(r_i) + \epsilon_i, \tag{3}$$

where, again because of the censorship, we do not observe all the values of T, but instead we observe the minimum variable (between the duration variable and the censoring one) Y. This model is very general because we do not assume any distribution for the duration or proportionality between the hazard functions.

In order to estimate model (3), we have to consider two main issues. On the one hand, the goodness of the fit and, on the other hand, the smoothness of the proposed function to model the effect of the covariate included in the nonparametric component. As for the

goodness of the fit, this is controlled through the sum of the weighted squared residuals using the Kaplan-Meier weights, calculated as in (2). Thus, using these weights, we take into account the existence of censored observations in the sample. As for the smoothness, we measure it in the usual way using the integral of the square of second derivatives. This can be handled by minimizing the following penalized weighted least squares expression

$$\sum_{i=1}^{n} W_{in}[\ln Y_{(i)} - X_i\beta - h(r_i)]^2 + \alpha \int \left[h^{"}(r)\right]^2 dr$$

The degree of smoothness is determined by α, the smoothing parameter. Large values of α produce smoother curves, while smaller values produce more wiggly curves. When α is close to zero, the penalty term becomes not relevant and the solution tends to an interpolating one. However, when α is large enough, the penalty term dominates and, thus, we obtain the weighted least squares solution.

Given α, the solution to the minimization problem described above is a smoothing cubic spline function and, using some properties of these functions, the previous expression can be rewritten as

$$(\ln Y - X\beta - Nh)^T W(\ln Y - X\beta - Nh) + \alpha h^T K h,$$

where h is the vector of values $h_j = h(r_j)$ for $j = 1, \ldots, d$, where d is the number of distinct values of the covariate R, X is the design matrix, N is the incidence matrix which assigns the respective value of the covariate R to each individual, W is a diagonal matrix with the Kaplan-Meier weights on its main diagonal, $\ln Y = (\ln Y_{(1)}, \ldots, \ln Y_{(n)})^T$ and K is a matrix obtained using the properties of the cubic spline function (for more details on splines see, e.g., Green and Silverman, 1994).

Taking derivatives in the expression above with respect to β and h, and reordering the terms leads us to obtain the next pair of simultaneous matrix equations

$$\begin{aligned} X^T W X\beta &= X^T W(\ln Y - Nh) \qquad (a) \\ (N^T W N + \alpha K)h &= N^T W(\ln Y - X\beta) \qquad (b) \end{aligned} \tag{4}$$

The estimations of β and h can be obtained iterating between equations 4(a) and 4(b), solving repeatedly for β and h, respectively, until convergence is achieved (i.e., using the backfitting algorithm).

The complete estimation process can be described by the following steps:

Previous steps

- **Step 1:** Separate the repeated values of the response variable
- **Step 2:** Calculate the Kaplan-Meier estimator $\hat{F}_n$ for the distribution function F
- **Step 3:** Calculate the Kaplan-Meier weights W_{in}
- **Step 4:** Put the observed response variable Y in increasing order
- **Step 5:** Put in adequate order, with the ordered Y, the covariates on the parametric component of the model (X) and the covariate of the nonparametric component (R)
- **Step 6:** Build the incidence matrix N

Backfitting

- **Step 7:** Obtain the initial estimated value of h $(\hat{h}_0)$ by applying ordinary least squares between $\ln Y$ and N
- **Step 8:** Substitute h by $\hat{h}_0$ in 4(a) and obtain $\hat{\beta}_0$ by weighted least squares using the Kaplan-Meier weights
- **Step 9:** Substitute β by $\hat{\beta}_0$ in 4(b) and obtain the new estimation of h $(\hat{h}_1)$ applying a natural cubic spline smoother to the difference $(\ln Y - X\beta)$

- **Step 10:** Go back to step 7 and continue until convergence is achieved

5. Inference Using Bootstrap Techniques

Once the estimation procedure is finished, we are interested in doing inference. In this paper, we carry out this analysis using computational methods, to be more precise, bootstrap resampling techniques. In order to do this, we have proposed a new procedure to generate the bootstrap resamples for the case of random censorship and a heterogeneous model. The bootstrap is relevant and its importance can be seen in the fact that it allows us to study the properties of the estimators even for small samples.

If we review the literature on the bootstrap with censored observation, we can basically find two different possibilities to obtain the bootstrap samples: proposed by Reid (1981) and Efron (1981), respectively.

The procedure proposed by Efron (1981) consists in estimating, by Kaplan-Meier, the distribution functions for the duration variable and for the censoring one, $\hat{F}_n$ and $\hat{G}_n$. Then, using these estimated distribution functions, we generate one sample for the duration variable, $t_1^*, \ldots, t_n^*$, and another one for the censoring variable, $c_1^*, \ldots, c_n^*$. Finally, we consider the following bootstrap resample:

$$y_i^* = \min\{t_i^*, c_i^*\}, \qquad \delta_i^* = \begin{cases} 1; & \text{if } t_i^* \leq c_i^* \\ 0; & \text{if } t_i^* > c_i^* \end{cases} .$$

On the other hand, the procedure proposed by Reid (1981) consists in estimating the Kaplan-Meier estimator for the distribution function of the duration variable $\hat{F}_n$ and, using this, generate the bootstrap resample. Akritas (1986) showed that the procedure proposed by Efron is better than the one considered by Reid.

However, these two resample generating methods were proposed to be applied in homogeneous models, that is, for models without covariates. In our case, we have covariates

because we want to estimate the effect of these covariates on the duration. Thus, the proposed resample procedures are not adequate for our case. However, the procedure of Efron can still be valid if we assume that the censoring variable follows the same regression model as the duration one. But, this assumption is very restrictive. In order to solve this problem, we propose a new procedure to generate the bootstrap samples for this sort of models. This procedure is very flexible because it does not assume any model for the relationship between the censoring variable and the covariates.

The complete procedure to obtain the bootstrap estimations can be described as follows:

- **Step 1:** Estimate model (3) following the proposal described in Section 4
- **Step 2:** Obtain the residuals of the previously estimated model:

$$\hat{\epsilon}_i = \ln Y_{(i)} - X_i\hat{\beta} - (N\hat{h}(r))_i; \quad \text{for} \quad i = 1, \ldots, n$$

- **Step 3:** Center the residuals
- **Step 4:** Obtain the bootstrap resample for the centered residuals $\epsilon_1^*, \ldots, \epsilon_n^*$
- **Step 5:** Generate the bootstrap sample for the variable of interest doing model-based boots-
  trap

$$\ln T_i^* = X_i\hat{\beta} + (N\hat{h}(r))_i + \epsilon_i^*; \quad \text{for} \quad i = 1, \ldots, n,$$

- **Step 6:** Generate a vector of Bernoulli variables δ^* where

$$P(\delta_i^* = 1 | \ln T_i^* = \ln t_i^*, X_i = x_i) = 1 - G(\ln t_i^{*-}), \quad \text{for} \quad i = 1, \ldots, n,$$

  and obtain the bootstrap indicator of censoring.

- **Step 7:** Estimate model (3), for the bootstrap sample, using the same estimation procedure as in Step 1. That is:

$$\min_{\beta,h} \sum_{i=1}^{n} W_{in}^{*} [\ln Y_{(i)}^{*} - X_i\beta - h(r_i)]^2 + \alpha \int \left[h^{"}(r)\right]^2 dr,$$

- **Step 8:** Go back to Step 4 and repeat the process M times (i.e., M bootstrap samples are obtained).

In Step 6, we should have obtained the bootstrap resample for the censoring variable and, then, comparing it with the bootstrap resample for the duration variable, obtain the minimum variable and the bootstrap censoring indicator. However, for the estimation process we only need the indicator of censoring and not the value of the censoring variable. Therefore, using Step 6 we can obtain the bootstrap indicator of censoring without assuming any relation between C and X, which is less restrictive than the proposal of Efron. In this step G denotes the distribution function of the censoring variable and, since it is unknown, we use its Kaplan-Meier estimator, $\hat{G}_n$. In Step 7, and for each bootstrap replication, we have to obtain the estimates using the procedure presented in Section 4. The value of M, in Step 8, depends on the objective of the study. If we want to estimate the distribution of the estimators or to obtain confidence intervals, we need a large value, at least $M = 1000$. However, if we are just interested in their standard deviations, far lower values are sufficient. For more details about bootstrap procedures see, e.g., Davison and Hinkley (1997) or Efron and Tibshirani (1993).

6. Application: A Flexible Model for Modelling the Survival Time in AIDS Patients

In this section, we present the application of the new methodology proposed to the data presented in Section 3. We estimate model (3) introducing in the parametric specification

all of the covariates except the period of diagnosis, which is introduced in a nonparametric term. Thus, we want to model the effect of this variable in a flexible way, to capture the effect of the AZT treatment and, maybe, some other effects.

In order to do this, we build a new variable which indicates the period of diagnosis of the illness for each patient. Thus, for patients diagnosed in the second quarter of 1984 this variable takes value one and, for patients diagnosed in the last quarter of 1990, it takes value twenty seven.

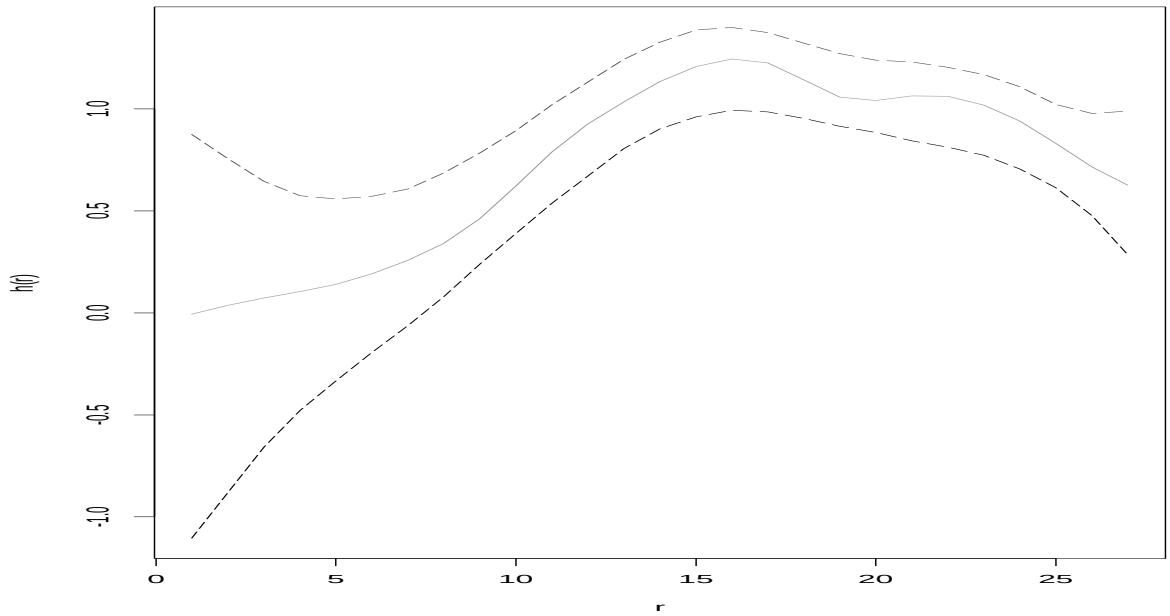
We estimate the model and, using the bootstrap techniques, calculate standard deviations (SDEV) and confidence intervals (lower limit, LL, and upper limit, UL). We can summarize the most relevant results obtained for our dataset: In Table 2, the results for the parametric component of the model and, in Figure 1, the ones for the nonparametric component.

Table 2: Estimates of β and 95% confidence intervals (Censored Partial Regression Model)

VARIABLE	COEF	SDEV	LL	UL
Constant	0.5266	0.4312	-0.3611	1.3357
Sex	0.0348	0.1425	-0.2501	0.3106
Disease1	-0.0154	0.2730	-0.5672	0.5080
Disease2	-0.0508	0.3339	-0.6882	0.6272
T-Sex	-0.1201	0.2520	-0.6020	0.3610
T-Drug	-0.0430	0.2174	-0.4722	0.3779
T-Blood	0.0702	0.2954	-0.4866	0.6971
T-Moth-child	0.3877	0.5577	-0.6330	1.5835
Age	-0.0177	0.0067	-0.0313	-0.0052

With regard to the covariates introduced in the parametric component, as in the previous analysis shown in Section 3, the age of the patient has a negative significant effect on his/her survival time. The rest of the covariates in the parametric component are not significant to explain the survival time of the patient.

Figure 1: Estimation and 95% confidence intervals for the function h



As for the estimation of the nonparametric component, using this model, we have the possibility of concluding that the effect of the period of diagnosis is not significant at the beginning of the illness and, as time goes by, it has a significant and positive effect with a clear acceleration, several quarters before the beginning of the administration of AZT (the administration of this treatment started in the middle of 1987, i.e. $r = 13$). This makes sense because patients whose diagnosis time was several quarters before starting the administration of AZT also receive this treatment. Therefore, we can say that the introduction of AZT has a positive effect on the survival, increasing the survival time of patients. This result agrees with the ones obtained in different works mentioned in Brookmeyer and Gail (1993) as, for example, Lemp et al. (1990). As a main conclusion, we can say that with this procedure we are able to detect the gradual effect of the administration of AZT.

Clearly, a dummy specification would not be adequate because the real effect changes progressively with time, and this cannot be captured using a dummy variable specification. In addition, with this model, we can evaluate the survival time for patients diagnosed in different periods and, thus, we can see the evolution of his/her survival time. In the first quarters, the effect of the period of diagnosis on the survival time is small and, as time goes by, this effect increases. This can be explained by the fact that AIDS was quite unknown at the beginning and, then, became widely known in our society. As a result, the disease was diagnosed earlier (i.e. as soon as it was developed by the patient). Then, we observe a strong acceleration, increasing the survival time because of the introduction of AZT, as pointed out above. Finally, we want to mention that the slight final drop is caused by the data, because the distance from the last quarters to the end of 1992 (when we finish the follow up of the patients) is not big enough to observe the complete durations in all the cases and, therefore, the maximum reachable duration is smaller when we are approximating the last quarters available in the sample. If we had extended the follow up period of the patients, this final drop would have not occurred.

7. Simulation Studies

The objective of this section is twofold. On the one hand, we want to verify the ability of the proposed estimation process in the censored partial regression model. On the other, we would like to analyze the effect of the censorship level on the estimation of the parametric and nonparametric components. In order to do this, the study has been carried out using different levels of censorship and different sample sizes.

We have generated a duration variable with log-normal probability distribution, following the model

$$\ln T = 2 + 1X_1 + 3X_2 + e^{\sin X_3} + \epsilon, \quad \text{with} \quad \epsilon \in N(0, \sigma = 0.5),$$

where $X_1 \in U[0,3]$ and $X_2 \in U[0,1]$ are the explanatory variables, which after being generated, are considered fixed. For the nonparametric component, we use the function $e^{\sin X_3}$, a function with two peaks and one valley, where X_3 has been obtained randomly generating equally likely integer values, between 0 and 10. We have chosen this complicated function to analyze the goodness of the fit of the proposed estimation process. For the censoring variable, we have considered a variable, independent from the duration and from the explanatory variables, and distributed as a uniform random variable (the interval for this uniform variable changes with the required level of censorship). We consider three levels of censorship, 10%, 20% and 33%, two different sample sizes, $n = 100$ and $n = 200$ and, for each combination, we generate 1000 samples.

The results for the parametric component are shown in Figures 2, 3, 4 and 5, presenting the box-plots of β estimated coefficients.

Figure 2: Estimates for β_1 ($n = 100$)

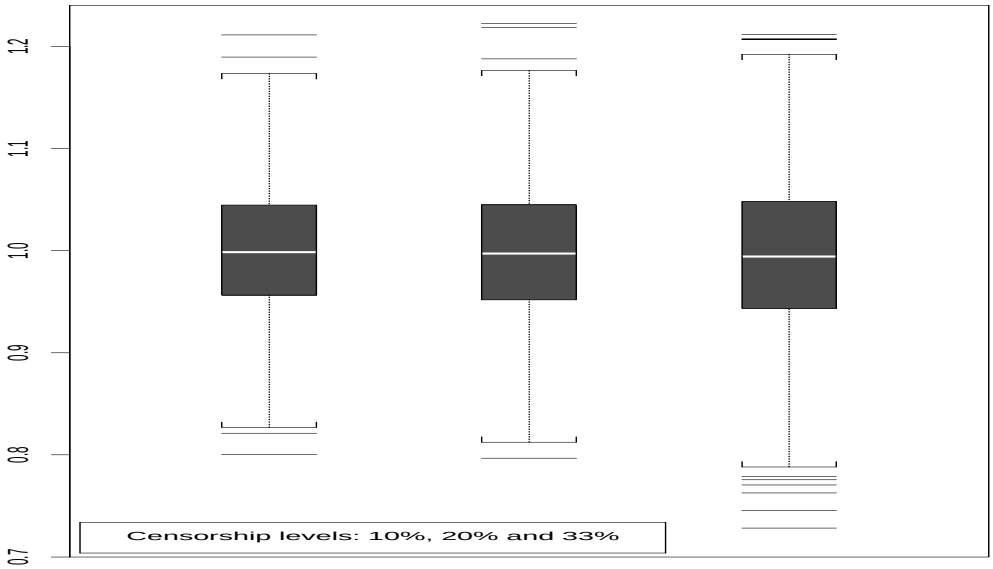


Figure 3: Estimates for β_1 ($n = 200$)

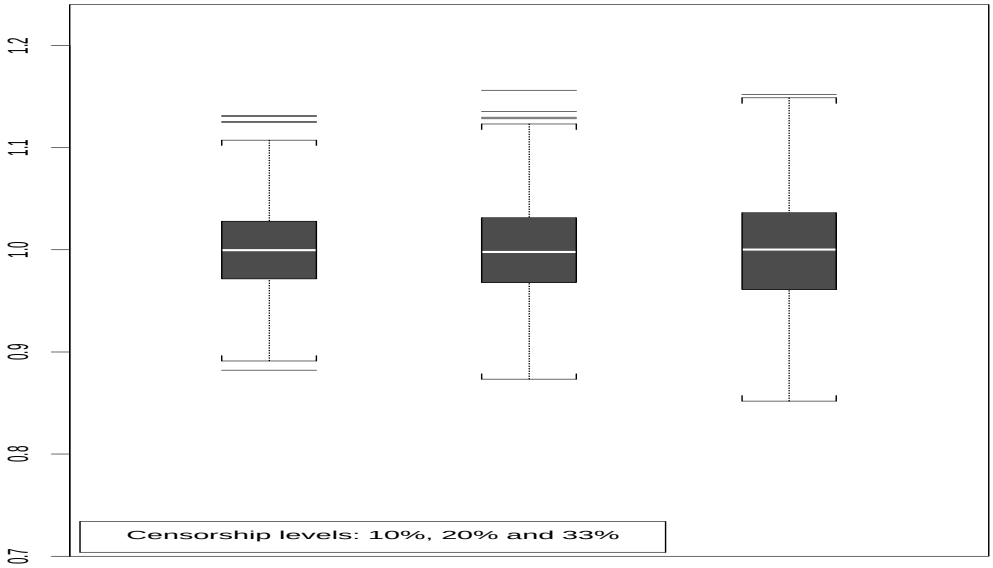


Figure 4: Estimates for β_2 ($n = 100$)

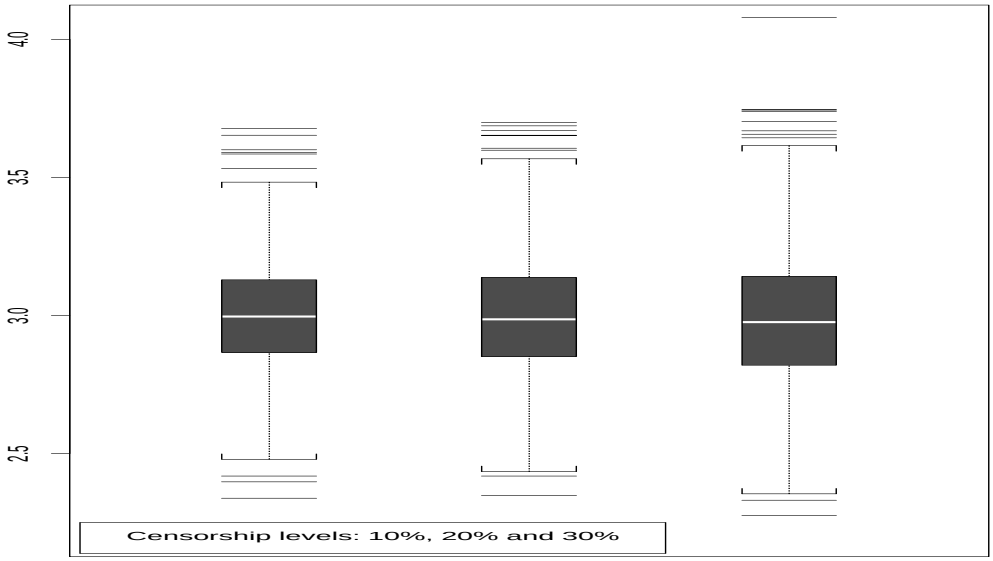
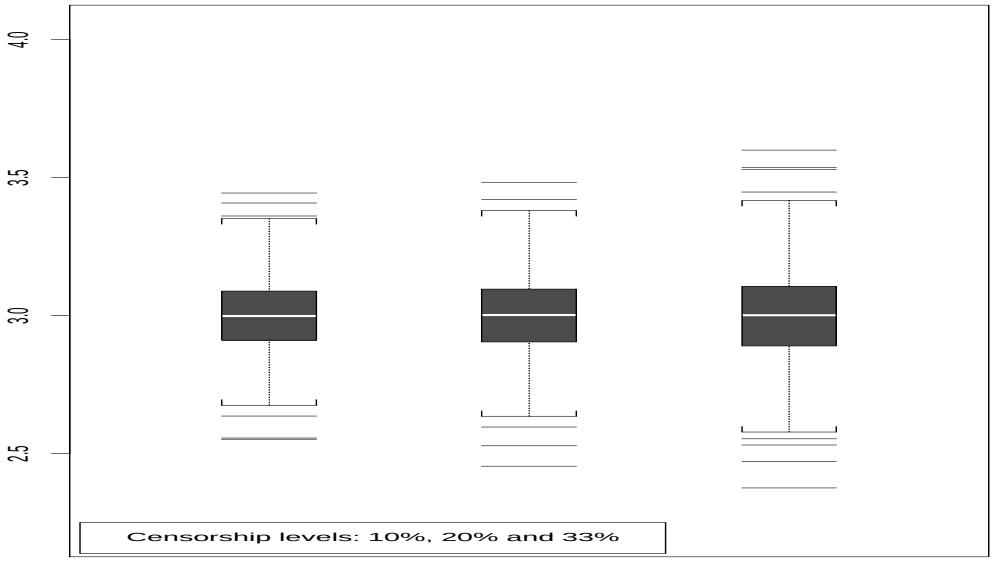


Figure 5: Estimates for β_2 ($n = 200$)



Figures 2 and 3 show the box-plots of the β coefficient associated to the variable X_1 for $n = 100$ and $n = 200$, respectively. In each figure, we have the results for the different censorship levels considered, on the left box we have a 10% censoring level, on the middle box, 20%, and on the right box, 33%. Figures 4 and 5 show the same information but for the coefficient associated to the variable X_2.

In order to do a better comparison, we present the estimated mean values and variances for $n = 100$ (Table 3) and for $n = 200$ (Table 4).

To summarize the results for the estimation of the nonparametric component, we use the following measurement error

$$\text{ME} = \frac{1}{11}\sum_{i=1}^{11}\left[f(x_{3i}) - \widehat{f}(x_{3i})\right]^2$$

Table 3: β_1 and β_2 coefficients estimation (n=100)

	β_1		β_2		
Censoring level	Mean	Variance	Mean	Variance	MSE (total)
10 %	0.9988	0.0039	2.9939	0.0393	0.0432
20 %	0.9977	0.0045	2.9935	0.0475	0.0521
33 %	0.9936	0.0061	2.9843	0.0625	0.0689

Table 4: β_1 and β_2 coefficients estimation (n=200)

	β_1		β_2		
Censoring level	Mean	Variance	Mean	Variance	MSE (total)
10 %	0.9998	0.0018	3.0003	0.0171	0.0188
20 %	0.9998	0.0022	2.9989	0.0195	0.0217
33 %	0.9982	0.0030	2.9968	0.0262	0.0292

After calculating these mean errors we present them in the box-plots (one for each censoring level) for $n = 100$ (Figure 6), and for $n = 200$ (Figure 7).

Figure 6: Mean errors for the nonparametric estimation (n=100)

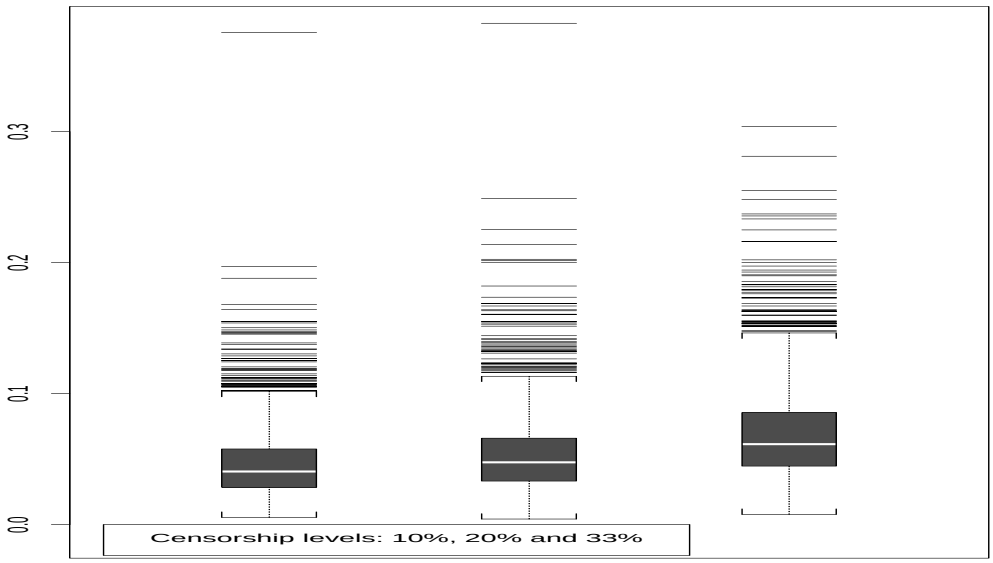
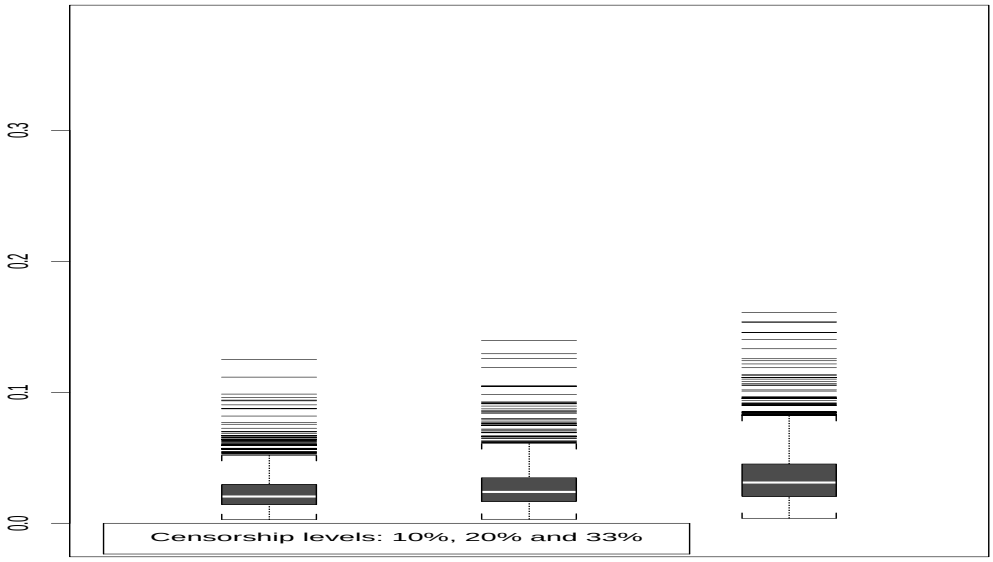


Figure 7: Mean errors for the nonparametric estimation (n=200)



If we take a look at the information shown above, we can see that the proposed model produces good estimates for both the parametric and nonparametric components.

As for the parametric component (Figures 2 to 5), we see that the median of the boxes is located at the real value of the coefficient. The effect of the censoring, as expected, tends to increase the variance of the estimations, reaching less precise estimates as the censoring level increases. In addition and, also as expected, the estimation is better as the sample size increases.

As for the nonparametric component the conclusions are similar when the censoring level increases, the boxes are wider and they move up. Again, if we increase the sample size the results improve substantially.

8. Conclusions

In this paper we have proposed a new methodology to analyze a response variable and the effect that some covariates have on it when we have censored samples. Our proposal is based on a model that does not need to assume any distribution for the duration variable, or the proportionality of the hazard functions for different individuals. In addition, we model directly the effect of the covariates on the duration, instead of the conditional probability to pass from one state (for example life) to a different one (for example death) at time t conditioned on the event of having stayed in that state until t. It also allows us to model situations where we do not know the functional form of the effect of one covariate on the response variable.

We use bootstrap techniques to make inference on the estimators for the censored partial regression model and, in order to do this, we have proposed a new bootstrap procedure to obtain the bootstrap samples for heterogeneous models with random censorship. This new procedure is a very general one because it does not assume any model for the relation between the censoring mechanism and the covariates.

The simulation study indicates that the proposed procedure to estimate the censored partial regression model produces good estimates for the parametric component and for the nonparametric one, even in the case of a complicated function, as the one used in the simulations.

We present an application of the proposed model with a real dataset where we analyze the survival of AIDS diagnosed patients, concluding that the age of the patient and the period of diagnosis are relevant factors to explain the survival time. We conclude indicating that the partial censored regression model could be used to study duration data in a flexible way in other different contexts such as, for example, Engineering or Economics.

References

M.G. Akritas, "Bootstrapping the Kaplan-Meier estimator," Journal of the American Statistical Association vol. 81 pp. 1032-1038, 1986.

R. Brookmeyer and M.H. Gail, AIDS Epidemiology a Quantitative Approach, Oxford University Press: Oxford, 1993.

D.R. Cox, "Regression models and life-tables," Journal of the Royal Statistical Society-Series B vol. 34 pp. 187-220, 1972.

D.R. Cox, "Partial likelihood," Biometrika vol. 62 pp. 269-276, 1975.

A.C. Davison and D.V. Hinkley, Bootstrap Methods and Their Application, Cambridge University Press: Cambridge, 1997.

B. Efron, "The two sample problem with censored data," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 4, Berkeley, 1967, pp. 831-853.

B. Efron, "Censored data and bootstrap," Journal of the American Statistical Association vol. 76 pp. 312-319, 1981.

B. Efron and R.J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall: New York, 1993.

P.J. Green and B.W. Silverman, Nonparametric Regression and Generalized Linear Models, Chapman and Hall: London, 1994.

E.L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," Journal of the American Statistical Association vol. 53 pp. 457-481, 1958.

J.F. Lawless, Statistical Models and Methods for Lifetime Data, John Wiley and Sons: New York, 1982.

G.P. Lemp, S.F. Payne and D. Neal, "Survival trends for patients with AIDS," Journal of the American Medical Association vol. 263 pp. 402-406, 1990.

J. Orbe, A. Fernández, and V. Nuñez-Antón, "Análisis de la supervivencia en enfermos de SIDA residentes en la Comunidad Autónoma del País Vasco y Navarra," Osasunkaria vol. 12 pp. 1-8, 1996.

N. Reid, "Estimating the median survival time," Biometrika vol. 68 pp. 601-608, 1981.

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

W. Stute, "Distributional convergence under random censorship when covariables are present," Scandinavian Journal of Statistics vol. 23 pp. 461-471, 1996a.

W. Stute, "The jackknife estimate of variance of a Kaplan-Meier integral," The Annals of Statistics vol. 24 pp. 2679-2704, 1996b.