Jesus Orbe; Vicente Núñez-Antón

Bias correction on censored least squares regression models

Kybernetika, Vol. 48 (2012), No. 5, 1045--1063

Persistent URL: http://dml.cz/dmlcz/143098

KYBERNETIKA — VOLUME *48* (2012), NUMBER 5, PAGES 1045–1063

BIAS CORRECTION ON CENSORED LEAST SQUARES REGRESSION MODELS

Jesus Orbe and Vicente Núñez-Antón

This paper proposes a bias reduction of the coefficients' estimator for linear regression models when observations are randomly censored and the error distribution is unknown. The proposed bias correction is applied to the weighted least squares estimator proposed by Stute [28] [W. Stute: Consistent estimation under random censorship when covariables are present. J. Multivariate Anal. *45* (1993), 89–103.], and it is based on model-based bootstrap resampling techniques that also allow us to work with censored data. Our bias-corrected estimator proposal is evaluated and its behavior assessed in simulation studies concluding that both the bias and the mean square error are reduced with the new proposal.

Keywords: bias, censoring, least squares, linear regression, Kaplan–Meier estimator

Classification: 62N01, 62F40

1. INTRODUCTION

In duration and survival data analysis we are usually unable to completely observe the variable of interest T, known as lifetime, failure time, duration or survival. Instead, we observe $Y = \min(T, C)$, where C represents the censoring variable. One of the most important objectives in this area, as in the case of non-censored data, is to be able to measure or estimate the effects the X covariates have on the censored variable under study, T. Researchers working on censored data analysis have treated this problem from two possible modelling approaches. The first one consists of modelling the covariates' effect on the hazard function or hazard rate $\lambda(t,x) = f(t,x)/[1 - F(t,x)]$, where $f(\cdot)$ and $F(\cdot)$ are, respectively, the probability density and distribution functions for the censored variable of interest T. Under this approach, the most commonly used model is the proportional hazards regression model (PH) proposed by Cox [6]:

$$\lambda(t,x) = \lambda_0(t)e^{x^T\beta},$$

where β is the vector of regression coefficients and $\lambda_0(t)$ is an unspecified function. The second approach consists of directly analyzing the effect the covariates have on the censored variable of interest T or on some transformation of it, such as for example, $\ln T$, with the use of a linear regression model:

$$t_i = x_i^T\beta + \epsilon_i$$

Under the second approach we have the accelerated failure time models (AFT). The most commonly used model is the proportional hazards regression model and the main reasons for its use are: (i) the possibility of estimating the covariates' effect without assuming any specific probability distribution for the censored variable T after maximizing the partial likelihood function (see [7]) and, (ii) the availability of this methodology in almost every statistical software package. However, this model is based on the proportional hazards assumption and this may not hold in general for some studies. Moreover, in a review of survival analyses in cancer journals, Altman et al. [2] reported that only five per cent of all studies using the Cox PH model attempted to verify the PH assumption. In addition, Cox indicated in Reid [24] that the accelerated failure time models are in many ways more appealing because of their quite direct physical interpretation. Besides, Stare et al. [27] claimed that results obtained from the Cox model fits are more difficult to explain to non-statisticians and provide less information than those obtained from linear regression fits. Furthermore, if the objective of the study is to be able to make predictions, then the linear regression modelling approach is clearly more appropriate than the hazard regression modelling approach.

Wei [36] and Stare et al. [27] pointed out that if there are no censored data, the most likely modelling approach to be used would be the linear regression modelling approach instead of the proportional hazards regression modelling approach. Therefore, the linear regression model could clearly be an interesting alternative to the proportional hazards model. However, its main disadvantage is that the usual estimation procedure for AFT models requires the assumption of some probability distribution for T and is based on maximizing the likelihood function taking into account the presence of censored observations. In addition, another drawback of the AFT modelling approach is that this probability distribution function is, in most cases, unknown to the practitioners. In order to solve this problem, several alternative approaches have been proposed in the literature for the estimation of the accelerated failure time models without the need to assume any probability distribution function for T. More specifically, rank-based methods for censored data have been studied in Tsiatis [35], Ritov [25], Lai and Ying [17] and Jin et al. [14], and least squares based methods for censored data have been investigated by Miller [20], Buckley and James [3], Koul et al. [16] and Stute [28].

In this paper, we will focus on least squares based methods, which, as we have just mentioned, have been approached from quite different perspectives. Our focus on these methods as alternative to the aforementioned ones is mainly motivated by the fact that, in our view, least squares methods are flexible enough to be able to handle censored data situations and, in addition, these methods are commonly and widely used, and well understood. More specifically, as will be described later in this section, we focus on Stute [28]'s approach because it is general enough, flexible, it does not require any iterative procedure for estimating the parameters, and, in addition, it can be easily extended to partial linear models (Orbe et al. [22]) or to nonlinear models (Stute [33]). Miller [20] used a least squares minimization procedure with a weighted sum of squares. These weights took into account the effect of the censored observations and were computed by estimating the distribution function of the error distribution based on the residuals of the linear regression using the Kaplan–Meier estimator (see Kaplan and Meier [15]). Thus, this methodology requires an iterative procedure and, in addition and for the consistency

of the estimators, it also requires that the censoring variable C have the same assumed regression model as that of the variable T. Buckley and James [3] proposed to replace each censored data point y_i by an estimate of the conditional expectation $E(T_i|T_i > y_i)$ based on the Kaplan–Meier estimator of the error distribution, which is estimated from the residuals of the linear regression, and then, they applied the usual least squares procedure. This methodology also requires an iterative procedure and, as the authors have pointed out, iterations may eventually settle down to oscillation between two values, although it seems that the values of the estimator proposed by Buckley and James are closer than those of the estimator advocated by Miller. Koul et al. [16], proposed to replace each observed data point y_i by the estimated value of $\delta_i y_i[1 - G(y_i)]^{-1}$, where $\delta_i = I(t_i \leq c_i)$, and G is the distribution function of the censoring variable, which could be estimated by using the Kaplan–Meier estimator, and then, they also applied the usual least squares procedure. The main advantage of this latter estimator is that it is not necessary to use an estimating iterative procedure. However, Koul et al. [16] suggested to truncate larger observations in the data. In addition, this method requires that the censoring variable be independent from the covariates in the model. Stute [28] proposed a weighted least squares estimator, where the weights considered the effect of the censoring data and were computed estimating the distribution function of the T variable based on the Kaplan–Meier weights of the observed variable Y. This estimator is quite simple to implement, it does not require any computational iteration scheme, it is consistent under minimal distributional assumptions (see [28]), it allows for random covariates, and it is easy to generalize to the multiple linear regression models case or to any other more complex models, such as for example, partial linear models (see Orbe et al. [22]) or nonlinear models (see Stute [33]).

There are several papers where the aforementioned proposals are compared. Miller and Halpern [21] concluded that the Buckley and James estimator proposal is more reliable than those by Miller and Koul et al. Heller and Simonoff [13] compared the Buckley and James estimator with those proposed by Chatterjee and McLeish [4], Leurgans [18] and Schmee and Hahn [26], and concluded that the Buckley and James estimator is clearly preferred. Stute [28] compared the Buckley and James estimator with his own proposal and the results indicated that his proposed estimator outperformed that of the Buckley and James estimator. Against this backdrop, this paper focuses on Stute's [28] approach and its main objective is to propose a bias-corrected estimator for Stute's proposal that can improve his results for small samples and also for situations where the bias is an important problem. As will be seen later in the paper, this new proposal not only reduces the bias but also the mean square error of the estimators.

The rest of the paper is organized as follows. Section 2 describes the estimation method proposed by Stute [28]. Our proposal for bias correction is introduced in Section 3. Section 4 assesses the performance of our proposal and compares it to that of previous ones through simulation studies. Section 5 provides some concluding remarks.

2. A CENSORED REGRESSION MODEL

We now briefly describe the methodology proposed by Stute [28]. 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

$$y_i = \min(t_i, c_i), \qquad \delta_i = \begin{cases} 1; & \text{if } t_i \le c_i \\ 0; & \text{if } t_i > c_i, \end{cases}$$

where $c_1, \ldots, c_n$ are the values for the censoring variable C from some unknown distribution function G, which is assumed to be independent of the duration variable T, and δ_i is the indicator for the observed failure. In addition, x_i represents the k-dimensional vector of covariates for the ith individual and, for identifiability reasons, Stute [28] assumed that $P(T \le C|T, X) = P(T \le C|T)$. As Stute [28] indicates, this condition states that given the time of failure, the covariates do not provide any further information as to whether censoring will take place or not. In addition, this condition is also satisfied if either C is independent of (T, X) or δ and X are independent conditionally on T (see, Stute [33]). Under these settings, we consider that the relation between the covariates and the duration is given by

$$\ln t_i = x_i^T \beta + \epsilon_i \quad \text{with} \quad \mathrm{E}[\epsilon_i | x_i] = 0, \tag{1}$$

where the ϵ_i's are assumed to be i.i.d. random variables. The estimator of β can be obtained by minimizing $\sum_{i=1}^n W_{in}[\ln y_{(i)} - x_i^T \beta]^2$, where x_i is the vector of covariates associated to the i-ordered Y-value, $y_{(1)} \le \ldots \le y_{(n)}$ are the ordered Y-values, and where ties between censored values or uncensored values are arbitrarily ordered, and ties between uncensored and censored times are treated as if the former precedes the latter. Thus, $y_{(i)}$ is the ith ordered value of the observed response variable Y and W_{in} is the Kaplan–Meier weight of the i-order statistic (i. e., the mass attached to the $y_{(i)}$ order statistic). These weights can be calculated as:

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

where $\hat{F}_n$ is a Kaplan–Meier estimator (see [15]) of the distribution function F, and $\hat{F}_n(y_{(0)}) \equiv 0$. These weights can be also calculated by using the redistribute to the right algorithm in Efron [9]. For the uncensored case, these weights take $1/n$ value and, thus, this proposal coincides with that of the ordinary least squares method. Thus, the estimator for β is given by

$$\hat{\beta} = (X^T W X)^{-1} (X^T W \ln Y), \tag{3}$$

where $\ln Y = (\ln y_{(1)}, \ldots, \ln y_{(n)})^T$, W is a diagonal matrix with the Kaplan–Meier weights on its main diagonal, and $X = [x_1, \ldots, x_n]^T$ is the design matrix or matrix of covariates. Model (1) can be considered within the class of accelerated failure time models. However, it allows for the estimation without assuming any distribution for the duration variable and, therefore, this model can be considered as an interesting alternative to the previously proposed ones.

The theoretical properties for this estimator for regression with censored data are provided under very general hypotheses. Stute [28] extended the nonparametric maximum likelihood estimator of the distribution function F proposed by Kaplan and Meier

[15], $\hat{F}_n(t)$, to a multivariate Kaplan–Meier estimator $\hat{F}_n^0(t,x)$ to be able to study the joint distribution of X and T. He proved a general strong law for Kaplan–Meier integrals, $\int \varphi d\hat{F}_n^0$, for a general function φ instead of for an indicator function, which is the usual in the literature on censored data estimation. Based on this idea, he proved the consistency of the new regression parameter estimator in the context of the censored linear regression model. Stute [31] proved the asymptotically normal distribution for the Kaplan–Meier integrals when covariates are present and Stute [32] provided a consistent jackknife estimate of the variance of a Kaplan–Meier integral.

3. BOOTSTRAP BIAS CORRECTION PROPOSAL

The issue of the resulting bias for Kaplan–Meier integrals has been previously investigated. In fact, for an indicator function, Gill [12] proved that $\hat{F}_n(t)$ is always biased downwards. Mauro [19] extended this result to a general Kaplan–Meier function, and Zhou [38] established a lower bound for this bias. Stute [29] derived an explicit formula for the bias and, in addition, he indicated that the bias of a Kaplan–Meier integral may be zero if there is no censoring, or it may decrease to zero at different rates depending on the φ function and censoring level in the right tails. Therefore, the objective of the reduction of the bias is an important task indeed. For the case of indicator functions, Chen et al. [5] proposed to modify the Kaplan–Meier estimator $\hat{F}_n(t)$, so that the indicator of the observed failure for the largest observed value $y_{(n)}$ takes value 1. That is, $\delta_n = 1$ regardless of whether the largest observed value is uncensored or not. Wellner [37] compared both estimators and concluded that the upward bias of the Chen et al.'s proposal was worse than the downward bias of the usual Kaplan–Meier estimator. Using the same bias reduction ideas, Stute [30] proposed a modification of the Kaplan–Meier estimator $\hat{F}_n(t)$ based on a bootstrap correction term of the cumulative hazard function. In order to be able to extend his proposal to a general Kaplan–Meier integral, Stute and Wang [34] presented the jackknife estimator of the bias of the Kaplan–Meier integral and a bias-corrected jackknife estimate of the Kaplan–Meier integral. More specifically, their proposal corrects the estimator of a Kaplan–Meier integral using the estimated jackknife bias of a Kaplan–Meier integral. As described in Stute and Wang [34], this estimated jackknife bias is obtained with the same ideas from the usual methodology for not censored cases, but replacing the empirical distribution function with the Kaplan–Meier estimator of the distribution function. They basically show that the proposed jackknife correction modifies the weight assigned to the largest observation, W_{nn}, by $W_{nn} + (n-1)\alpha_n/n$, where α_n is the weight associated to the largest observation in the Kaplan–Meier estimator, computed for the whole sample that does not include the one before the last ordered observation.

All of the aforementioned research on the bias of Kaplan–Meier integrals has studied situations where the covariates are not present in the model. In this paper we concentrate on the reduction of the bias for the estimator proposed by Stute for the regression coefficients. We are not aware of any scientific study or attempt that has tried to correct the bias in those situations, where the estimator for the regression coefficients is biased. In this way, as can be seen in (3), this estimator is calculated using the Kaplan–Meier integrals, where the elements of the matrices $(X^T W X)$ and $(X^T W \ln Y)$ are Kaplan–Meier integrals of different φ functions. Therefore, we try to reduce the bias

using and extending the ideas in Stute and Wang [34] for the estimator of the regression coefficients and, in addition and most importantly, given that it is the main objective in this paper, we also put forward a different approach that consists in estimating the bias by using bootstrap techniques, and then building the bias-corrected estimator. 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.

If we review the literature on bootstrap with censored observations, we can basically find two different approaches to obtain the bootstrap samples, Reid [23] and Efron [10]. The procedure proposed by Efron [10] consists in estimating, by Kaplan–Meier, the distribution functions for the duration and the censoring variables, $\hat{F}_n$ and $\hat{G}_n$. Then, using these estimated distribution functions, he generated one sample for the duration variable, $t_1^*, \ldots, t_n^*$, and another for the censoring variable, $c_1^*, \ldots, c_n^*$. Finally, and as described in Section 2, he considered the following bootstrap resample:

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

As an alternative, the procedure proposed by Reid [23] consists in estimating, also by Kaplan–Meier, the distribution function for the duration variable $\hat{F}_n$ and, in generating the bootstrap resample using this estimator. Akritas [1] concluded that the procedure proposed by Efron is better than the one considered by Reid. 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 in the model because our modelling objectives focus on the possibility of estimating the effect these covariates have on the duration variable. For these specific settings, we have two possible approaches: (i) resampling cases approach, i. e., resampling from (y_i, δ_i, x_i); or (ii) a model-based resampling approach. If there is a belief that the model is correctly specified and we have a fixed design, the model-based approach is more appropriate (for more details, see, e. g., Davison and Hinkley [8], or Efron and Tibshirani [11]). Therefore and given that we have a fixed design and the process generates the sample for a given assumed model, we propose a new model-based approach bootstrap procedure that can handle the presence of covariates in the duration model. It is a regression model-based bootstrap that does not require the assumption of any distribution for the error term, and that also allows us to work with censored data. In addition, it is a general procedure that can consider different censoring schemes to generate the bootstrap samples, which is the main motivation for the proposal included in this paper. Moreover, the process uses Stute's identifiability condition described in Section 2, which is a weaker condition than that of assuming independence between C and (T, X). Finally, our proposal is very flexible because it does not need to assume any model or probability distribution for the censoring variable. We now describe the steps required to obtain the proposed bootstrap bias-corrected estimations:

- **Step 1:** Following the proposal described in Section 2, estimate model (1).
- **Step 2:** Obtain the residuals for the aforementioned estimated model:

$$\hat{\epsilon}_i = \ln y_{(i)} - x_i^T \hat{\beta}, \quad \text{for} \quad i = 1, \ldots, n.$$

- **Step 3:** Using the residuals in Step 2 and the indicator of the observed failure; that is, $(\hat{\epsilon}_i, \delta_i)$, compute the distribution function of the residuals by using the Kaplan–Meier estimator and, thus, obtain the bootstrap resample for the errors: $\epsilon_1^*, \ldots, \epsilon_n^*$.

- **Step 4:** Generate the bootstrap sample for the variable of interest by doing model-based bootstrap. That is,

$$\ln t_i^* = x_i^T \hat{\beta} + \epsilon_i^*; \quad \text{for} \quad i = 1, \ldots, n.$$

- **Step 5:** Generate a vector of Bernoulli variables δ_i^*, where

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

  and obtain the bootstrap indicator for the observed failure. Here, $\hat{G}$ denotes the Kaplan–Meier estimator of the distribution function for the censoring variable, where $\hat{G}$ is computed as in the case of the Kaplan–Meier estimator of the distribution function for the duration variable but changing the indicator for the observed failure, so that it now assigns value one to a censored observation.

- **Step 6:** Generate the censoring variable. If $\ln T_i^* = \ln t_i^*$ and $\delta_i^* = 1$, C_i^* is taken from $\hat{G}$ restricted to the interval $[t_i^*, +\infty)$, and if $\ln T_i^* = \ln t_i^*$ and $\delta_i^* = 0$, C_i^* is taken from $\hat{G}$ restricted to the interval $[0, t_i^*)$.

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

$$\min_{\beta} \sum_{i=1}^{n} W_{in}^* [\ln y_{(i)}^* - x_i^T \beta]^2.$$

- **Step 8:** Go back to Step 3 and repeat the process M times (i. e., M bootstrap samples are obtained). This will generate the corresponding M bootstrap estimates for the parameter β, $\hat{\beta}^{*(1)}, \ldots, \hat{\beta}^{*(M)}$.

- **Step 9:** Using the M bootstrap estimates for parameter β in Step 8, obtain the bias bootstrap estimate:

$$\widehat{\text{bias}} = \frac{\sum_{m=1}^{M} \hat{\beta}^{*(m)}}{M} - \hat{\beta}.$$

- **Step 10:** Finally, obtain the bootstrap bias-corrected estimator, defined as

$$\hat{\beta}_{c2} = \hat{\beta} - \widehat{\text{bias}} = 2\hat{\beta} - \frac{\sum_{m=1}^{M} \hat{\beta}^{*(m)}}{M}.$$

Note that, in Step 5, we use the aforementioned identifiability condition assumed by Stute [28]; that is, $P(T \leq C | T, X) = P(T \leq C | T)$. Moreover, by jointly using Steps 5

and 6 above, we generate the variables (X, T, C), where the conditions that T and C are independent and, in addition, that $P(T \leq C|T, X) = P(T \leq C|T)$ hold. We would also like to mention that not every joint distribution of the vector (X, T, C) satisfying these aforementioned conditions can be generated by the proposed algorithm. More specifically the algorithm may not reflect the possible dependence between C and X. Of course, these two conditions are also satisfied if C is independent of (T, X). Finally, in Step 7, as in Section 2, we obtain the bootstrap resample of observed durations as

$$\ln y_i^* = \begin{cases} \ln t_i^*; & \text{if } \delta_i^* = 1 \\ \ln c_i^*; & \text{if } \delta_i^* = 0 \end{cases}$$

The value of M in Step 8 depends on the objective of the study. If we wish to estimate the distribution of the estimators or to obtain confidence intervals, we need a large value of at least $M = 1000$. However, if we are just interested in obtaining their standard deviations or bias, far lower values are sufficient. For more details about bootstrap procedures see, e. g., Davison and Hinkley [8] or Efron and Tibshirani [11].

4. SIMULATION STUDY

We have conducted simulation studies to be able to assess the performance of the proposed bias-corrected estimator. The values of the duration variable of interest T were generated from the model:

$$\ln T = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \tag{4}$$

where X_1 and X_2 follow a uniform distribution on the interval $(0, 5)$, $\beta_0 = \beta_1 = \beta_2 = 1$, and ϵ is normally distributed with mean 0 and as standard deviation taking different values $\sigma = \{1, 0.75, 0.5\}$. The values for the censoring variable C were generated by using different uniform distribution functions. We consider three possible censoring levels: 15 %, 30 %, and 50 %. Table 1 presents the estimated bias results for Stute's estimator (i. e., $\hat{\beta}$), for the bias-corrected jackknife estimator using the idea presented in Stute and Wang [34] (i. e., $\hat{\beta}_{c1}$), and for the bootstrap bias-corrected estimator proposed here (i. e., $\hat{\beta}_{c2}$). Results reported in Table 1 correspond to a sample of size $n = 40$ based on 1000 simulated datasets and using $M = 199$ bootstrap replicates in each generated dataset. A brief summary of the more relevant results follows:

- As can be seen in the results reported in Table 1, the resulting bias for the new proposal, $\hat{\beta}_{c2}$, is the smallest one when compared to the estimator without bias correction or to the jackknife bias-corrected estimator for all of the estimated regression parameters, for all σ values and censoring levels. The bias reduction observed for the new proposal is larger for large censoring levels and σ values.

- The univariate mean square error (mse) for each one of the estimated parameters is smaller for the bootstrap bias-corrected proposal in all cases considered here.

- If we analyze the global estimation performance, using as an indicator the multivariate mean square error, we can observe in Table 2 that, for all the cases

(a): Estimates with censoring level 50 %.

		β_0			β_1			β_2		
σ		bias	var	mse	bias	var	mse	bias	var	mse
1	$\hat{\beta}$	0.4388	0.3196	0.5122	-0.1789	0.0332	0.0652	-0.1769	0.0370	0.0683
	$\hat{\beta}_{c1}$	0.4374	0.3224	0.5137	-0.1781	0.0340	0.0657	-0.1758	0.0380	0.0689
	$\hat{\beta}_{c2}$	-0.0493	0.3485	0.3509	-0.0543	0.0441	0.0470	-0.0497	0.0466	0.0491
0.75	$\hat{\beta}$	0.3071	0.1900	0.2843	-0.1180	0.0203	0.0342	-0.1163	0.0241	0.0376
	$\hat{\beta}_{c1}$	0.3053	0.1913	0.2846	-0.1172	0.0207	0.0344	-0.1154	0.0246	0.0379
	$\hat{\beta}_{c2}$	-0.0345	0.1950	0.1962	-0.0293	0.0248	0.0257	-0.0244	0.0281	0.0287
0.5	$\hat{\beta}$	0.1696	0.0984	0.1272	-0.0616	0.0102	0.0140	-0.0635	0.0130	0.0170
	$\hat{\beta}_{c1}$	0.1685	0.0994	0.1278	-0.0612	0.0103	0.0141	-0.0630	0.0132	0.0172
	$\hat{\beta}_{c2}$	-0.0192	0.0933	0.0937	-0.0121	0.0115	0.0116	-0.0119	0.0136	0.0137

(b): Estimates with censoring level 30 %.

		β_0			β_1			β_2		
σ		bias	var	mse	bias	var	mse	bias	var	mse
1	$\hat{\beta}$	0.2611	0.2368	0.3050	-0.0855	0.0217	0.0290	-0.0944	0.0206	0.0295
	$\hat{\beta}_{c1}$	0.2508	0.2421	0.3050	-0.0822	0.0228	0.0295	-0.0903	0.0216	0.0298
	$\hat{\beta}_{c2}$	-0.0221	0.2499	0.2504	-0.0227	0.0246	0.0252	-0.0218	0.0247	0.0251
0.75	$\hat{\beta}$	0.1563	0.1393	0.1637	-0.0470	0.0126	0.0149	-0.0575	0.0125	0.0158
	$\hat{\beta}_{c1}$	0.1477	0.1445	0.1663	-0.0440	0.0136	0.0155	-0.0543	0.0133	0.0162
	$\hat{\beta}_{c2}$	-0.0145	0.1375	0.1377	-0.0101	0.0135	0.0136	-0.0127	0.0139	0.0141
0.5	$\hat{\beta}$	0.0740	0.0637	0.0692	-0.0204	0.0058	0.0062	-0.0280	0.0058	0.0066
	$\hat{\beta}_{c1}$	0.0708	0.0656	0.0706	-0.0194	0.0060	0.0064	-0.0268	0.0061	0.0069
	$\hat{\beta}_{c2}$	-0.0060	0.0623	0.0623	-0.0039	0.0061	0.0061	-0.0060	0.0063	0.0063

(c): Estimates with censoring level 15 %.

		β_0			β_1			β_2		
σ		bias	var	mse	bias	var	mse	bias	var	mse
1	$\hat{\beta}$	0.1336	0.2010	0.2189	-0.0402	0.0174	0.0190	-0.0441	0.0165	0.0185
	$\hat{\beta}_{c1}$	0.1213	0.2079	0.2227	-0.0371	0.0181	0.0195	-0.0391	0.0179	0.0195
	$\hat{\beta}_{c2}$	-0.0104	0.2021	0.2022	-0.0089	0.0184	0.0184	-0.0082	0.0178	0.0179
0.75	$\hat{\beta}$	0.0713	0.1179	0.1230	-0.0201	0.0100	0.0104	-0.0240	0.0095	0.0101
	$\hat{\beta}_{c1}$	0.0623	0.1217	0.1256	-0.0174	0.0105	0.0108	-0.0210	0.0100	0.0105
	$\hat{\beta}_{c2}$	-0.0077	0.1176	0.1177	-0.0032	0.0103	0.0103	-0.0038	0.0100	0.0100
0.5	$\hat{\beta}$	0.0250	0.0525	0.0532	-0.0077	0.0044	0.0044	-0.0081	0.0042	0.0043
	$\hat{\beta}_{c1}$	0.0190	0.0545	0.0548	-0.0062	0.0045	0.0046	-0.0060	0.0045	0.0045
	$\hat{\beta}_{c2}$	-0.0084	0.0517	0.0518	-0.0003	0.0044	0.0044	0.0003	0.0043	0.0043

Tab. 1. Estimated biases (bias), variances (var) and mean square errors (mse) for the coefficients without bias correction versus the two bias-corrected proposals for different values of $\sigma = \{1, 0.75, 0.5\}$ and different censoring levels, 50 %, 30 % and 15 %.

considered here, the proposed bootstrap bias-corrected estimator, $\hat{\beta}_{c2}$, presents the smallest multivariate mean square error. Here, the multivariate mean square error is defined as the sum of the individual univariate mean square errors for the three estimators being considered in the regression model (i. e., $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$) for each one of the proposed bias-corrected estimators.

σ	%C	50	30	15
1	$\hat{\beta}$	0.6457	0.3635	0.2564
1	$\hat{\beta}_{c1}$	0.6483	0.3643	0.2617
1	$\hat{\beta}_{c2}$	0.4470	0.3007	0.2385
0.75	$\hat{\beta}$	0.3561	0.1944	0.1435
0.75	$\hat{\beta}_{c1}$	0.3569	0.1980	0.1469
0.75	$\hat{\beta}_{c2}$	0.2506	0.1654	0.1380
0.5	$\hat{\beta}$	0.1582	0.0820	0.0619
0.5	$\hat{\beta}_{c1}$	0.1591	0.0839	0.0639
0.5	$\hat{\beta}_{c2}$	0.1190	0.0747	0.0605

Tab. 2. Multivariate mean square errors for estimations without bias correction versus the two bias-corrected proposals for different values of $\sigma = \{1, 0.75, 0.5\}$ and different censoring levels, 50 %, 30 % and 15 %.

$(\sigma, \%C)$	50	30	15
1	1.445	1.209	1.075
0.75	1.421	1.175	1.040
0.5	1.329	1.098	1.023

Tab. 3. Ratio of multivariate mean square errors for estimations without bias correction (i. e., $\hat{\beta}$) versus the proposed bootstrap bias-corrected procedure (i. e., $\hat{\beta}_{c2}$) for different values of $\sigma = \{1, 0.75, 0.5\}$ and different censoring levels, 50 %, 30 % and 15 %.

- In relation to the previous comment, Table 3 shows that the advantage of using the proposed bootstrap bias-corrected estimator is greater when the censoring level increases and/or the value of σ is larger. Thus, for example, if we consider the case of a 50 % censoring level and $\sigma = 1$, the loss when using the non-corrected version compared to the bootstrap bias-corrected version results in an increment of 44.5 % on the multivariate mean square error.

- In addition, for the three estimators considered here, when the value for the parameter σ decreases, the biases and variances corresponding to each coefficient decrease and, thus, both the univariate and multivariate mean square errors also decrease. This variance effect is quite clear because by changing σ we increase or decrease the model's variability. As suggested by Heller and Simonoff [13], the effect over the bias could be induced by the asymmetric effect of increasing the σ parameter on the censoring of the response variable. Thus, a larger positive ϵ_i error in model (1) corresponds to a probably larger and censored $\ln t_i$ response, while a larger negative ϵ_i error corresponds to a probably uncensored $\ln t_i$ response.

(a): Estimates with censoring level 50 %.

σ		bias	β_0 var	mse	bias	β_1 var	mse	bias	β_2 var	mse
1	$\hat{\beta}$	0.5303	0.3178	0.5990	-0.2387	0.0277	0.0847	-0.2354	0.0328	0.0882
	$\hat{\beta}_{c1}$	0.5302	0.3178	0.5989	-0.2387	0.0277	0.0847	-0.2353	0.0329	0.0882
	$\hat{\beta}_{c2}$	-0.1819	0.3288	0.3618	-0.0613	0.0373	0.0411	-0.0552	0.0429	0.0460
0.75	$\hat{\beta}$	0.3756	0.1984	0.3395	-0.1629	0.0179	0.0444	-0.1521	0.0226	0.0457
	$\hat{\beta}_{c1}$	0.3754	0.1985	0.3394	-0.1629	0.0179	0.0444	-0.1520	0.0227	0.0458
	$\hat{\beta}_{c2}$	-0.1325	0.1924	0.2099	-0.0308	0.0216	0.0225	-0.0203	0.0253	0.0257
0.5	$\hat{\beta}$	0.1830	0.1004	0.1339	-0.0811	0.0095	0.0160	-0.0663	0.0124	0.0168
	$\hat{\beta}_{c1}$	0.1829	0.1005	0.1339	-0.0810	0.0095	0.0160	-0.0663	0.0124	0.0168
	$\hat{\beta}_{c2}$	-0.0681	0.0895	0.0941	-0.0097	0.0100	0.0101	-0.0009	0.0115	0.0115

(b): Estimates with censoring level 30 %.

σ		bias	β_0 var	mse	bias	β_1 var	mse	bias	β_2 var	mse
1	$\hat{\beta}$	0.3766	0.2226	0.3644	-0.1267	0.0177	0.0337	-0.1422	0.0174	0.0377
	$\hat{\beta}_{c1}$	0.3764	0.2230	0.3647	-0.1266	0.0177	0.0337	-0.1419	0.0176	0.0377
	$\hat{\beta}_{c2}$	-0.0927	0.2412	0.2498	-0.0246	0.0225	0.0231	-0.0272	0.0226	0.0234
0.75	$\hat{\beta}$	0.2455	0.1272	0.1875	-0.0773	0.0103	0.0162	-0.0916	0.0100	0.0184
	$\hat{\beta}_{c1}$	0.2449	0.1273	0.1873	-0.0770	0.0103	0.0162	-0.0914	0.0100	0.0184
	$\hat{\beta}_{c2}$	-0.0657	0.1353	0.1396	-0.0117	0.0124	0.0125	-0.0122	0.0121	0.0123
0.5	$\hat{\beta}$	0.1135	0.0580	0.0709	-0.0348	0.0047	0.0059	-0.0421	0.0047	0.0065
	$\hat{\beta}_{c1}$	0.1132	0.0580	0.0708	-0.0348	0.0047	0.0059	-0.0419	0.0047	0.0065
	$\hat{\beta}_{c2}$	-0.0347	0.0599	0.0611	-0.0037	0.0054	0.0054	-0.0025	0.0053	0.0053

(c): Estimates with censoring level 15 %.

σ		bias	β_0 var	mse	bias	β_1 var	mse	bias	β_2 var	mse
1	$\hat{\beta}$	0.1849	0.1970	0.2312	-0.0545	0.0163	0.0193	-0.0636	0.0142	0.0182
	$\hat{\beta}_{c1}$	0.1841	0.1970	0.2309	-0.0542	0.0164	0.0194	-0.0633	0.0142	0.0182
	$\hat{\beta}_{c2}$	-0.0292	0.2085	0.2094	-0.0096	0.0179	0.0180	-0.0107	0.0167	0.0168
0.75	$\hat{\beta}$	0.1009	0.1104	0.1205	-0.0281	0.0092	0.0099	-0.0344	0.0083	0.0095
	$\hat{\beta}_{c1}$	0.1003	0.1105	0.1206	-0.0279	0.0092	0.0100	-0.0342	0.0083	0.0095
	$\hat{\beta}_{c2}$	-0.0169	0.1133	0.1136	-0.0034	0.0096	0.0096	-0.0047	0.0092	0.0092
0.5	$\hat{\beta}$	0.0430	0.0495	0.0513	-0.0123	0.0041	0.0043	-0.0139	0.0038	0.0040
	$\hat{\beta}_{c1}$	0.0424	0.0495	0.0513	-0.0122	0.0041	0.0043	-0.0137	0.0038	0.0040
	$\hat{\beta}_{c2}$	-0.0117	0.0510	0.0512	-0.0008	0.0042	0.0042	-0.0003	0.0041	0.0041

Tab. 4. Estimated biases (bias), variances (var) and mean square errors (mse) for the coefficients without bias correction versus the two bias-corrected proposals for different values of $\sigma = \{1, 0.75, 0.5\}$ and different censoring levels, 50 %, 30 % and 15 % for a constant censoring scheme.

- We can also try to reduce the width of the support for the censoring up to a minimum value, so that we are artificially changing the case of random censoring to that of not random censoring (i. e., when the censoring distribution is constant). In practice, this is a very common and realistic situation. We can then observe the

results reported in Table 4, where the bias reduction and the improvement of the multivariate mean square errors still hold for the proposed bootstrap bias-corrected estimator, $\hat{\beta}_{c2}$.

- It is also important to mention that the width of the support for the censoring variable has an effect on the resulting bias as well. More specifically, a smaller width of the support of the censoring variable results in a larger bias for all of the three estimators being compared here (see Tables 1 to 4). This result is consistent with the one obtained in Heller and Simonoff [13] for other estimation methods. In addition, this result can be easily motivated by the fact that a smaller width in the support of the censoring variable implies that the upper limit for the censoring is also smaller. As a consequence, we lose information about the duration variable on the right tail of its distribution probability. In this problematic but common situation, the advantage of using the proposed bootstrap bias-corrected estimator is much greater because the bias reduction, and also that of the multivariate mean square error, is larger (see Table 3).

- Finally, we have to point out that, as expected, as the censoring level decreases, the mean square errors in each estimator also decrease.

As can be seen from the aforementioned conclusions, the bootstrap bias-corrected proposal behaves reasonably well, reducing both the estimator's bias and mean square error. In addition, we have studied the performance of the bootstrap bias-corrected proposal considering different alternative scenarios that may be of interest to researchers in the area and possibly common in real data set situations. We now describe these new simulations and the conclusions we have obtained from them.

- We start by considering a situation where we have correlated discrete and continuous covariates. Thus, we consider the situation where the duration variable T is generated as in equation (4), but where the covariate X_1 follows a binomial probability distribution with parameters $p = 1/2$ and $n = 5$, and the covariate X_2 is generated in such a way that it is correlated with X_1. That is, $X_2 = X_1 + v$, where v is assumed to be normally distributed with mean 0 and standard deviation 1. The error term, ϵ, has been generated as in the previous simulation study. That is, ϵ is normally distributed with mean 0 and standard deviation taking different values $\sigma = \{1, 0.75, 0.5\}$. The censoring variable C has been generated by using different uniform distributions, so that we can consider three possible censoring levels: 15 %, 30 %, and 50 %. Table 5 presents the results for the estimators' biases and mean square errors for all of the cases considered here.

- We have also considered other alternative probability distributions for the duration variable T. That is, distributions different from the log-normal distribution for T, which is equivalent to considering distributions different from the normal for the ϵ term. In this way, we have studied the case of an exponential regression model, which assumes a constant hazard function and, in addition, two different Weibull regression models, where the assumption of a constant hazard function is relaxed. In order to do this, the duration variable T is now generated from the model:

$$\ln T = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma\epsilon, \tag{5}$$

where X_1 and X_2 follow a uniform distribution on the interval $(0, 5)$, $\beta_0 = \beta_1 = \beta_2 = 1$, but now ϵ follows an extreme value distribution, and, in addition, γ takes on two possible different values, 1.5 and 0.5 for the Weibull regression model, and value 1 for the exponential regression model. The corresponding supports for the censoring variable C have been appropriately modified so that three possible censoring levels are obtained: 50 %, 30 % and 15 %. Results for the estimators' biases and mean square errors are reported in Table 6, for the exponential regression model, and in Table 7, for the Weibull regression models.

- Finally, we have also considered two other censoring scheme distributions. The first one studies the case generated by model (4) but where the censoring variable is now generated from a normal probability distribution with changing mean value, so that three possible censoring levels are obtained: 50 %, 30 % and 15 %, and with standard deviation taking value 1. The second one considers the situation of a censoring distribution that depends on one of the covariates in the model. In this way, we use model (4) with a censoring variable generated as $\ln C = a + bX_2 + u$, where u is normally distributed with mean 0 and standard deviation σ. Different values of a and b are used so that the aforementioned censoring levels are also obtained. Tables 8 and 9 present the results for the estimators' biases and mean square errors under these two settings.

- As a brief summary of the results obtained under these alternative settings and reported in Tables 5 to 9, we can see that the bias for the bootstrap bias-corrected proposal are, in general, smaller than those of the estimator without the bias correction. As for the mean square error, results in some of the new simulations are very similar for the bootstrap bias-corrected proposal and for the estimator without the bias correction and, thus, the gain in terms of mean square error is, in those cases, quite modest.

5. CONCLUSIONS

It is very common to study a problem where the variable under study is not completely observed because of censoring. That is, we really observe either the actual duration, if the observation is not censored, or only know that the duration is larger that a given value, if the observation is censored. Under these settings, if the appropriate estimator is not used, we can obtain estimators that are seriously biased. For example, in a situation where the censoring dominates the right tail of the response's probability distribution, a serious problem of bias can arise, which could in turn produce misleading conclusions. However, this would also be a smaller problem when the data are analyzed by regression models based on hazard functions than in models based on transformations of the censored outcome. The goal of this paper is to put forward a new proposal that tries to reduce the bias of the coefficients' estimator for linear regression models. Thus, we propose a new bootstrap bias-corrected estimator for the regression coefficient estimators for censored data originally proposed by Stute [28]. This new procedure is appropriate to be applied in heterogeneous censored models (i. e., with the presence of covariates). In addition, it is also very general because it does not assume any model or

(a): Estimates with censoring level 50 %

σ		bias	β_0 var	mse	bias	β_1 var	mse	bias	β_2 var	mse
1	$\hat{\beta}$	0.2842	0.2239	0.3047	-0.2264	0.1221	0.1734	-0.1257	0.0764	0.0922
	$\hat{\beta}_{c2}$	-0.1163	0.2198	0.2333	-0.0683	0.1451	0.1497	-0.0286	0.0858	0.0866
0.75	$\hat{\beta}$	0.1995	0.1332	0.1730	-0.1554	0.0719	0.0961	-0.0801	0.0433	0.0497
	$\hat{\beta}_{c2}$	-0.0645	0.1243	0.1285	-0.0387	0.0762	0.0777	-0.0200	0.0463	0.0467
0.5	$\hat{\beta}$	0.1074	0.0635	0.0750	-0.0823	0.0352	0.0420	-0.0409	0.0199	0.0215
	$\hat{\beta}_{c2}$	-0.0265	0.0568	0.0575	-0.0161	0.0345	0.0347	-0.0113	0.0206	0.0208

(b): Estimates with censoring level 30 %.

σ		bias	β_0 var	mse	bias	β_1 var	mse	bias	β_2 var	mse
1	$\hat{\beta}$	0.2142	0.1673	0.2131	-0.1307	0.0764	0.0935	-0.0900	0.0584	0.0665
	$\hat{\beta}_{c2}$	-0.1135	0.1685	0.1814	-0.0311	0.0887	0.0897	-0.0150	0.0629	0.0631
0.75	$\hat{\beta}$	0.1253	0.0969	0.1126	-0.0727	0.0451	0.0504	-0.0536	0.0326	0.0355
	$\hat{\beta}_{c2}$	-0.0609	0.0962	0.0999	-0.0090	0.0505	0.0505	-0.0117	0.0335	0.0336
0.5	$\hat{\beta}$	0.0623	0.0436	0.0475	-0.0367	0.0212	0.0225	-0.0243	0.0155	0.0161
	$\hat{\beta}_{c2}$	-0.0215	0.0421	0.0425	-0.0061	0.0224	0.0224	-0.0058	0.0157	0.0157

(c): Estimates with censoring level 15 %.

σ		bias	β_0 var	mse	bias	β_1 var	mse	bias	β_2 var	mse
1	$\hat{\beta}$	0.1698	0.1474	0.1762	-0.0623	0.0642	0.0681	-0.0865	0.0486	0.0560
	$\hat{\beta}_{c2}$	-0.0814	0.1464	0.1530	-0.0108	0.0715	0.0716	-0.0169	0.0516	0.0519
0.75	$\hat{\beta}$	0.0995	0.0875	0.0974	-0.0338	0.0359	0.0371	-0.0518	0.0287	0.0314
	$\hat{\beta}_{c2}$	-0.0410	0.0819	0.0835	-0.0048	0.0387	0.0387	-0.0088	0.0303	0.0304
0.5	$\hat{\beta}$	0.0421	0.0393	0.0411	-0.0096	0.0163	0.0164	-0.0273	0.0133	0.0140
	$\hat{\beta}_{c2}$	-0.0172	0.0368	0.0371	0.0024	0.0170	0.0170	-0.0075	0.0137	0.0137

Tab. 5. Estimated biases (bias), variances (var) and mean square errors (mse) for the coefficients without bias correction versus the bootstrap bias-corrected proposal considering correlated discrete and continuous covariates for different values of $\sigma = \{1, 0.75, 0.5\}$ and different censoring levels, 50 %, 30 % and 15 %.

probability distribution for the censoring variable and, because of its flexibility, it can consider different censoring schemes to generate the bootstrap samples.

We have studied the behavior of the proposed bootstrap bias-corrected estimator for some very general simulation settings. The bias for this estimator has proved to be the smallest for all the situations considered here. Our proposal has been compared to the non-corrected version of the estimator and with another bias-corrected estimator that is based on the ideas originally proposed by Stute and Wang [34] for a general Kaplan–Meier integral without covariates. The bias is reduced when using the bootstrap bias-corrected estimator proposal. In addition, the multivariate mean square errors for our new proposal are always smaller than those of the non-corrected estimator.

Comparisons for bias reduction and for the ratio of the multivariate mean square

C		bias	β_0 var	mse	bias	β_1 var	mse	bias	β_2 var	mse
50 %	$\hat{\beta}$	0.5479	0.5313	0.8315	-0.2274	0.0568	0.1085	-0.1983	0.0654	0.1048
	$\hat{\beta}_{c2}$	-0.0737	0.6817	0.6872	-0.0750	0.0850	0.0906	-0.0475	0.0948	0.0970
30 %	$\hat{\beta}$	0.4244	0.4075	0.5876	-0.1388	0.0353	0.0546	-0.1391	0.0448	0.0641
	$\hat{\beta}_{c2}$	-0.0533	0.4643	0.4671	-0.0303	0.0432	0.0441	-0.0326	0.0559	0.0569
15 %	$\hat{\beta}$	0.2588	0.2904	0.3574	-0.0753	0.0208	0.0265	-0.0728	0.0290	0.0343
	$\hat{\beta}_{c2}$	-0.0163	0.3299	0.3301	-0.0198	0.0243	0.0247	-0.0177	0.0325	0.0328

Tab. 6. Estimated biases (bias), variances (var) and mean square errors (mse) for the coefficients without bias correction versus the bootstrap bias-corrected proposal for an exponential regression model for different censoring levels, 50 %, 30 % and 15 %.

(a): Weibull regression model with $\gamma = 1.5$

C		bias	β_0 var	mse	bias	β_1 var	mse	bias	β_2 var	mse
50 %	$\hat{\beta}$	0.9028	0.8767	1.6918	-0.3235	0.0911	0.1958	-0.4093	0.1362	0.3037
	$\hat{\beta}_{c2}$	0.1490	1.5687	1.5910	-0.1579	0.1491	0.1740	-0.2210	0.2468	0.2956
30 %	$\hat{\beta}$	0.6068	0.7619	1.1302	-0.1860	0.0745	0.1091	-0.2439	0.0949	0.1544
	$\hat{\beta}_{c2}$	0.0395	1.0906	1.0922	-0.0679	0.1033	0.1079	-0.0959	0.1441	0.1533
15 %	$\hat{\beta}$	0.3300	0.6399	0.7488	-0.0854	0.0621	0.0694	-0.1121	0.0672	0.0798
	$\hat{\beta}_{c2}$	0.0394	0.7306	0.7321	-0.0313	0.0692	0.0702	-0.0340	0.0821	0.0832

(b): Weibull regression model with $\gamma = 0.5$

C		bias	β_0 var	mse	bias	β_1 var	mse	bias	β_2 var	mse
50 %	$\hat{\beta}$	0.2320	0.1374	0.1912	-0.0716	0.0132	0.0183	-0.0982	0.0231	0.0327
	$\hat{\beta}_{c2}$	0.0058	0.1526	0.1526	-0.0239	0.0156	0.0162	-0.0283	0.0285	0.0293
30 %	$\hat{\beta}$	0.0909	0.1010	0.1093	-0.0244	0.0088	0.0094	-0.0336	0.0122	0.0133
	$\hat{\beta}_{c2}$	-0.0066	0.1042	0.1042	-0.0058	0.0093	0.0093	-0.0028	0.0136	0.0136
15 %	$\hat{\beta}$	0.0583	0.0807	0.0841	-0.0150	0.0070	0.0072	-0.0170	0.0082	0.0085
	$\hat{\beta}_{c2}$	0.0059	0.0787	0.0788	-0.0058	0.0070	0.0070	-0.0026	0.0083	0.0083

Tab. 7. Estimated biases (bias), variances (var) and mean square errors (mse) for the coefficients without bias correction versus the bootstrap bias-corrected proposal for a Weibull regression model with different values of the γ parameter, $\gamma = \{1.5, 0.5\}$, and different censoring levels, 50 %, 30 % and 15 %.

errors show that the improvement when using the new proposal increases when the censoring level increases, the width of the support for the censoring variable decreases, or the σ parameter increases. In the first two cases, the effect of the censoring becomes stronger in the right tail of the probability distribution and, therefore, we lose information about the variable under study, resulting in a larger bias. In the last case, the increment of the bias could be explained by the aforementioned asymmetric effect of increasing the σ parameter on the censoring.

(a): Estimates with censoring level 50 %.

		β_0			β_1			β_2		
σ		bias	var	mse	bias	var	mse	bias	var	mse
1	$\hat{\beta}$	0.5468	0.5615	0.8605	-0.2027	0.0476	0.0887	-0.1712	0.0483	0.0776
	$\hat{\beta}_{c2}$	0.1213	0.6269	0.6417	-0.0952	0.0591	0.0682	-0.0770	0.0589	0.0648
0.75	$\hat{\beta}$	0.3730	0.3533	0.4924	-0.1345	0.0324	0.0505	-0.1119	0.0307	0.0432
	$\hat{\beta}_{c2}$	0.0709	0.3591	0.3642	-0.0557	0.0369	0.0400	-0.0442	0.0345	0.0365
0.5	$\hat{\beta}$	0.1884	0.1737	0.2092	-0.0659	0.0152	0.0195	-0.0551	0.0157	0.0187
	$\hat{\beta}_{c2}$	0.0273	0.1665	0.1672	-0.0219	0.0163	0.0167	-0.0178	0.0167	0.0170

(b): Estimates with censoring level 30 %.

		β_0			β_1			β_2		
σ		bias	var	mse	bias	var	mse	bias	var	mse
1	$\hat{\beta}$	0.3288	0.3993	0.5073	-0.1015	0.0268	0.0371	-0.0912	0.0312	0.0395
	$\hat{\beta}_{c2}$	0.0358	0.4208	0.4221	-0.0332	0.0296	0.0307	-0.0320	0.0352	0.0362
0.75	$\hat{\beta}$	0.2043	0.2339	0.2756	-0.0604	0.0164	0.0200	-0.0546	0.0178	0.0208
	$\hat{\beta}_{c2}$	0.0217	0.2267	0.2272	-0.0176	0.0171	0.0174	-0.0166	0.0187	0.0190
0.5	$\hat{\beta}$	0.0997	0.1114	0.1213	-0.0272	0.0073	0.0080	-0.0274	0.0083	0.0091
	$\hat{\beta}_{c2}$	0.0083	0.1062	0.1063	-0.0058	0.0074	0.0074	-0.0076	0.0086	0.0087

(c): Estimates with censoring level 15 % .

		β_0			β_1			β_2		
σ		bias	var	mse	bias	var	mse	bias	var	mse
1	$\hat{\beta}$	0.1752	0.3250	0.3557	-0.0489	0.0186	0.0210	-0.0454	0.0224	0.0245
	$\hat{\beta}_{c2}$	0.0195	0.3278	0.3281	-0.0145	0.0195	0.0197	-0.0157	0.0236	0.0238
0.75	$\hat{\beta}$	0.1042	0.1826	0.1935	-0.0288	0.0109	0.0117	-0.0259	0.0122	0.0129
	$\hat{\beta}_{c2}$	0.0121	0.1828	0.1830	-0.0085	0.0111	0.0112	-0.0084	0.0128	0.0129
0.5	$\hat{\beta}$	0.0509	0.0821	0.0847	-0.0132	0.0047	0.0049	-0.0130	0.0055	0.0056
	$\hat{\beta}_{c2}$	0.0065	0.0813	0.0814	-0.0034	0.0048	0.0048	-0.0044	0.0056	0.0056

Tab. 8. Estimated biases (bias), variances (var) and mean square errors (mse) for the coefficients without bias correction versus the bootstrap bias-corrected proposal considering a normally distributed censoring variable for different values of $\sigma = \{1, 0.75, 0.5\}$ and different censoring levels, 50 %, 30 % and 15 %.

Finally, the results reported here support the fact that our bias-correction estimator proposal is useful because it substantially reduces the bias, also reducing the multivariate mean square errors. Moreover, in situations where the non-corrected estimator is unbiased or approximately unbiased, our proposal is also unbiased (i. e., the reduction of bias is very small in absolute value), and the multivariate mean square errors are quite similar.

ACKNOWLEDGEMENTS

This work was supported by Ministerio de Ciencia e Innovación, FEDER, the Department of Education of the Basque Government (UPV/EHU Econometrics Research Group) and Universidad del País Vasco UPV/EHU under research grants ECO2010-15332, MTM2010-14913,

(a): Estimates with censoring level 50 %.

			β_0			β_1			β_2	
σ		bias	var	mse	bias	var	mse	bias	var	mse
1	$\hat{\beta}$	0.2645	0.2740	0.3439	-0.2026	0.0348	0.0759	-0.0654	0.0288	0.0330
	$\hat{\beta}_{c2}$	-0.1316	0.2808	0.2981	-0.1331	0.0412	0.0589	0.0179	0.0327	0.0330
0.75	$\hat{\beta}$	0.2244	0.1703	0.2207	-0.1664	0.0234	0.0511	-0.0478	0.0164	0.0186
	$\hat{\beta}_{c2}$	-0.0726	0.1566	0.1619	-0.1132	0.0263	0.0391	0.0141	0.0177	0.0179
0.5	$\hat{\beta}$	0.1217	0.0851	0.1000	-0.1046	0.0129	0.0239	-0.0160	0.0079	0.0081
	$\hat{\beta}_{c2}$	-0.0227	0.0779	0.0784	-0.0753	0.0145	0.0201	0.0151	0.0082	0.0085

(b): Estimates with censoring level 30 %.

			β_0			β_1			β_2	
σ		bias	var	mse	bias	var	mse	bias	var	mse
1	$\hat{\beta}$	0.2256	0.2219	0.2729	-0.1272	0.0181	0.0343	-0.0512	0.0197	0.0223
	$\hat{\beta}_{c2}$	-0.0837	0.2194	0.2264	-0.0771	0.0194	0.0253	0.0083	0.0212	0.0213
0.75	$\hat{\beta}$	0.1989	0.1316	0.1711	-0.0999	0.0105	0.0205	-0.0466	0.0115	0.0137
	$\hat{\beta}_{c2}$	-0.0435	0.1238	0.1257	-0.0616	0.0109	0.0147	-0.0013	0.0119	0.0119
0.5	$\hat{\beta}$	0.1426	0.0631	0.0834	-0.0638	0.0048	0.0088	-0.0372	0.0053	0.0067
	$\hat{\beta}_{c2}$	-0.0147	0.0580	0.0582	-0.0396	0.0047	0.0063	-0.0084	0.0053	0.0054

(c): Estimates with censoring level 15 %.

			β_0			β_1			β_2	
σ		bias	var	mse	bias	var	mse	bias	var	mse
1	$\hat{\beta}$	0.1278	0.1742	0.1905	-0.0576	0.0118	0.0152	-0.0315	0.0140	0.0150
	$\hat{\beta}_{c2}$	-0.0466	0.1829	0.1850	-0.0295	0.0124	0.0133	0.0012	0.0144	0.0144
0.75	$\hat{\beta}$	0.0996	0.0999	0.1099	-0.0367	0.0068	0.0081	-0.0275	0.0076	0.0083
	$\hat{\beta}_{c2}$	-0.0250	0.1014	0.1020	-0.0165	0.0070	0.0072	-0.0037	0.0077	0.0077
0.5	$\hat{\beta}$	0.0635	0.0460	0.0500	-0.0183	0.0031	0.0034	-0.0196	0.0034	0.0037
	$\hat{\beta}_{c2}$	-0.0099	0.0455	0.0456	-0.0060	0.0031	0.0032	-0.0054	0.0034	0.0034

Tab. 9. Estimated biases (bias), variances (var) and mean square errors (mse) for the coefficients without bias correction versus the bootstrap bias-corrected proposal considering a censoring variable that depends on the covariates, for different values of $\sigma = \{1, 0.75, 0.5\}$ and different censoring levels, 50%, 30 % and 15 %.

IT-334-07 and UFI11/03. The authors also wish to thank two anonymous referees for providing thoughtful comments and suggestions which have led to substantial improvement of the presentation of the material in this paper.

(Received June 30, 2011)

REFERENCES

[1] M. G. Akritas: Bootstrapping the Kaplan–Meier estimator. J. Amer. Statist. Assoc. *81* (1986), 1032–1038.

[2] D. G. Altman, B. L. De Stavola, S. B. Love, and K. A. Stepniewska: Review of survival analyses published in cancer journals. British J. Cancer. *72* (1985), 511–518.

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

[4] S. Chatterjee and D. L. McLeish: Fitting linear regression models to censored data by least squares and maximum likelihood methods. Comm. Statist. Theory Methods *15* (1986), 3227–3243.

[5] Y. Y. Chen, M. Hollander, and N. A. Langberg: Small sample results for the Kaplan–Meier estimator. J. Amer. Statist. Assoc. *77* (1982), 141–144.

[6] D. R. Cox: Regression models and life-tables. J. R. Stat. Soc. Ser. B. *34* (1972), 187–220.

[7] D. R. Cox: Partial likelihood. Biometrika *62* (1975), 269–276.

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

[9] B. Efron: The two sample problem with censored data. In: Proc. 5th Berkeley Symposium *4* (1967), pp. 831–853.

[10] B. Efron: Censored data and the bootstrap. J. Amer. Statist. Assoc. *76* (1981), 312–319.

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

[12] R. D. Gill: Censoring and Stochastics Integrals. Math. Centre Tracts 124, Math. Centrum, Amsterdam 1980.

[13] G. Heller and J. S. Simonoff: A comparison of estimators for regression with a censored response variable. Biometrika *77* (1990), 515–520.

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

[15] E. L. Kaplan and P. Meier: Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. *53* (1958), 457–481.

[16] H. Koul, V. Susarla, and J. Van-Ryzin: Regression analysis with randomly right-censored data. Ann. Statist. *9* (1981), 1279–1288.

[17] T. L. Lai and Z. Ying: Linear rank statistics in regression analysis with censored or truncated data. J. Multivariate Anal. *40* (1992), 13–45.

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

[19] D. Mauro: A combinatoric approach to the Kaplan–Meier estimator. Ann. Statist. *13* (1985), 142–149.

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

[21] R. G. Miller and J. Halpern: Regression with censored data. Biometrika *69* (1982), 521–531.

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

[23] N. Reid: Estimating the median survival time. Biometrika *68* (1981), 601–608.

[24] N. Reid: A conversation with Sir David Cox. Statist. Sci. *9* (1994), 439–455.

[25] Y. Ritov: Estimation in a linear regression model with censored data. Ann. Statist. *18* (1990), 303–328.

[26] J. Schmee and G. J. Hahn: A simple method for regression analysis with censored data (with discussion). Technometrics *21* (1979), 417–434.

[27] J. Stare, F. Heinzl, and F. Harrel: On the use of Buckley and James least squares regression for survival data. In: New Approaches in Applied Statistics (A. Ferligoj and A. Mrvar, eds.), Metodološki zvezki 16, Ljubljana: Eslovenia, 2000, pp. 125–134.

[28] W. Stute: Consistent estimation under random censorship when covariables are present. J. Multivariate Anal. *45* (1993), 89–103.

[29] W. Stute: The bias of Kaplan–Meier integrals. Scand. J. Stat. *21* (1994), 475–484.

[30] W. Stute: Improved estimation under random censorship. Comm. Statist. Theory Methods *23* (1994), 2671–2682.

[31] W. Stute: Distributional convergence under random censorship when covariables are present. Scand. J. Stat. *23* (1996), 461–471.

[32] W. Stute: The jackknife estimate of variance of a Kaplan–Meier integral. Ann. Statist. *24* (1996), 2679–2704.

[33] W. Stute: Nonlinear censored regression. Statist. Sinica *9* (1999), 1089–1102.

[34] W. Stute and J. L. Wang: The jackknife estimate of a Kaplan–Meier integral. Biometrika *81* (1994), 602–606.

[35] A. A. Tsiatis: Estimating regression parameters using linear rank tests for censored data. Ann. Statist. *18* (1990), 354–372.

[36] L. J. Wei: The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat. Med. *11* (1992), 1871–1879.

[37] J. A. Wellner: A heavy censoring limit theorem for the product limit estimator. Ann. Statist. *13* (1985), 150–162.

[38] M. Zhou: Two-sided bias bound of the Kaplan–Meier estimator. Probab. Theory and Related Fields *79* (1988), 165–173.

Jesus Orbe, Departamento de Econometría y Estadística, Universidad del País Vasco UPV/EHU, Avenida Lehendakari Aguirre 83, E-48015 Bilbao. Spain.
e-mail: jesus.orbe@ehu.es

Vicente Núñez-Antón, Departamento de Econometría y Estadística, Universidad del País Vasco UPV/EHU, Avenida Lehendakari Aguirre 83, E-48015 Bilbao. Spain.
e-mail: vicente.nunezanton@ehu.es