Estimating regression parameters in linear regression model with a censored response variable

Jesus Orbe and Vicente Núñez-Antón

Abstract—In this work we study the effect of several covariates X on a censored response variable T with unknown probability distribution. In this context, most of the studies in the literature can be located in two possible general classes of regression models: models that study the effect the covariates have on the hazard function; and models that study the effect the covariates have on the censored response variable. Proposals in this paper are in the second class of models and, more specifically, on least squares based model approach. Thus, using the bootstrap estimate of the bias, we try to improve the estimation of the regression parameters by reducing their bias, for small sample sizes. Simulation results presented in the paper show that, for reasonable sample sizes and censoring levels, the bias is always smaller for the new proposals.

Keywords—censored response variable, regression, bias.

I. INTRODUCTION

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) Censored observations because 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 appropriate. Thus, we have to consider alternative more appropriate distributions such as, for example, the Weibull, exponential or Gamma distributions.

As a result, traditional methods applied in standard problems in Statistics can not be used. In order to solve this issue and taking into account the special characteristics of this type 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 available covariates being considered to explain T. There are two big classes of regression models that analyze the dependence between X and T. The proportional hazards (PH) models proposed by Cox [1] and the accelerated failure time (AFT) models (see, e.g., Lawless [2]).

In the Cox model, we have that the hazard function is modelled as

Authors are with Dept. Econometría y Estadística, Facultad de Ciencias Económicas y Empresariales, Universidad del País Vasco/Euskal Herriko Unibertsitatea, Avda. Lehendakari Agirre 83, E-48015 Bilbao, Spain (email: jesus.orbe@ehu.es).

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

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 using 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 by using the partial likelihood function proposed in Cox [3].

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 as a model that considers a direct relation between the duration and the covariates. That is,

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

In general, the estimation of this model is carried out by 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 one of an uncensored observation is given by the density function. The problem is that, in most cases, this probability distribution is unknown for the practitioners. In order to solve this problem, several approaches have been proposed for the estimation on the accelerated failure time model, where it is not required to assume any probability distribution for the response variable. Thus, rank-based methods for censored data have been proposed by Tsiatis [4], Lai and Ying [5] or Jin et al. [6], and least squares based methods for censored data have been investigated, for example, by Miller [7], Buckley and James [8] or Stute [9].

In this paper, we concentrate on least squares based methods and, more specifically, on the methodology proposed in [9]. Stute presents a weighted least square estimator, where the

weights take into account the effect of the censored data and are computed by estimating the distribution function of the T variable, based on the Kaplan-Meier weights of the observed variable Y. This estimator is easy to implement, it does not require any iteration scheme, it is consistent under minimal distributional assumptions (see [9]), it allows for random co-variates, and it can be easily generalized to the multiple linear regression model case or other more complicated models, such as, for example, partial linear models (see Orbe et al. [10]) or nonlinear models (see Stute [11]). We focus our attention on the aforementioned approach. Our main objective is to propose an improvement to it by presenting a bias correction alternative methodology for small sample sizes, where, as will be seen later in the paper, the proposal not only reduces the bias, but also reduces the mean square error for the estimators.

The rest of the paper is organized as follows. In Section II, we present a flexible alternative methodology for the classic AFT model and provide some general details about its estimation procedure. In addition, we also present a proposal to improve the estimators' bias by proposing a bias-corrected version of it, where the bias is estimated with the use of bootstrap resampling techniques. Section III provides some simulation results to study the behavior of the proposed estimator, and Section IV presents some final conclusions and recommendations.

II. METHODOLOGY

We use a method similar to the one proposed in [9], in which a new estimator, assuming very general hypotheses, was obtained using weighted least squares. In order to 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 assumed to be 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

$$T_i = X_i\beta + \epsilon_i \tag{1}$$

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 computed 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 (Kaplan and Meier [12]) of the distribution function F. In this way, the estimator for β is given by

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

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. Stute [9] studied the consistency of this estimator, and Stute [13] its asymptotic normal distribution. 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 and, in addition, it does not require the assumption of proportional hazard functions.

In this paper, our main concern is to reduce the bias of the aforementioned estimator for the regression coefficients. As can be seen in (3), this estimator is computed by using Kaplan-Meier integrals because the elements of the $X^T W X$ and $X^T W Y$ matrices are indeed Kaplan-Meier integrals of different functions. The bias of the Kaplan-Meier integrals has been previously studied (see, e.g., Gill [14], Mauro [15] or Stute [16]). Here, we wish to estimate the bias for the estimator of the regression coefficient $\hat{\beta}$ and, then, compute the bias-corrected estimator

$$\hat{\beta}_c = \hat{\beta} - \widehat{BIAS}$$

The estimation of the bias is based on bootstrap resampling techniques. In order to do this, we have proposed a new methodology to generate the bootstrap resamples for the case of random censorship and a heterogeneous model. The itemized procedure to obtain the bootstrap replications can be described as follows:

- Estimate model (1) following the proposal described in this section and obtain the residuals of the previously estimated model. That is,

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

- Using these residuals, obtain the bootstrap resample for the errors: $\epsilon_1^*, \ldots, \epsilon_n^*$
- Generate the bootstrap sample for the variable of interest, by doing model-based bootstrap. That is,

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

- Generate a vector of Bernoulli variables δ^*, where

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

and obtain the bootstrap indicator for the censoring variable. G denotes the distribution function for the censoring variable and, since it is unknown, we use its Kaplan-Meier estimator, $\hat{G}_n$.
- Generate the censoring variable. That is, if $T^* = t^*$ and $\delta^* = 1$, C^* is taken from $\hat{G}_n$ restricted to $[t^*, \infty)$ interval, whereas, if $T^* = t^*$ and $\delta^* = 0$, C^* is taken from $\hat{G}_n$ restricted to $[0, t^*)$ interval.
- Estimate model (1), for the bootstrap sample, by using the same original estimation procedure. That is:

TABLE I
ESTIMATED BIASES (BIAS) AND VARIANCES (VAR) FOR THE ESTIMATED
COEFFICIENTS WITHOUT BIAS CORRECTION ($\hat{\beta}$), VERSUS THE
BIAS-CORRECTED PROPOSAL ESTIMATOR ($\hat{\beta}_c$) FOR DIFFERENT VALUES OF
$\sigma = \{1, 0.75, 0.5\}$ AND 50% CENSORING LEVEL.

σ		β_0 BIAS	β_0 VAR	β_1 BIAS	β_1 VAR	β_2 BIAS	β_2 VAR
1	$\hat{\beta}$	0.5303	0.317	-0.2387	0.027	-0.2354	0.032
	$\hat{\beta}_c$	-0.1819	0.328	-0.0613	0.037	-0.0552	0.042
0.75	$\hat{\beta}$	0.3756	0.198	-0.1629	0.017	-0.1521	0.022
	$\hat{\beta}_c$	-0.1325	0.192	-0.0308	0.021	-0.0203	0.025
0.5	$\hat{\beta}$	0.1830	0.100	-0.0811	0.009	-0.0663	0.012
	$\hat{\beta}_c$	-0.0681	0.089	-0.0097	0.010	-0.0009	0.011

TABLE II
ESTIMATED BIASES (BIAS) AND VARIANCES (VAR) FOR THE ESTIMATED
COEFFICIENTS WITHOUT BIAS CORRECTION ($\hat{\beta}$), VERSUS THE
BIAS-CORRECTED PROPOSAL ESTIMATOR ($\hat{\beta}_c$) FOR DIFFERENT VALUES OF
$\sigma = \{1, 0.75, 0.5\}$ AND 30% CENSORING LEVEL.

σ		β_0 BIAS	β_0 VAR	β_1 BIAS	β_1 VAR	β_2 BIAS	β_2 VAR
1	$\hat{\beta}$	0.3766	0.222	-0.1267	0.017	-0.1422	0.017
	$\hat{\beta}_c$	-0.0927	0.241	-0.0246	0.022	-0.0272	0.022
0.75	$\hat{\beta}$	0.2455	0.127	-0.0773	0.010	-0.0916	0.010
	$\hat{\beta}_c$	-0.0657	0.135	-0.0117	0.012	-0.0122	0.012
0.5	$\hat{\beta}$	0.1135	0.058	-0.0348	0.004	-0.0421	0.004
	$\hat{\beta}_c$	-0.0347	0.059	-0.0037	0.005	-0.0025	0.005

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

- Go back to the second step and repeat the process M times (i.e., M bootstrap samples are obtained).

At the end of this procedure, we obtain M bootstrap replications for the aforementioned $\hat{\beta}$ estimated parameter. That is, $\hat{\beta}^{*1}, \ldots, \hat{\beta}^{*M}$ and, therefore, we can derive the bias bootstrap estimate as

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

Finally, we obtain the bootstrap bias-corrected estimator by using

$$\hat{\beta}_c = \hat{\beta} - \widehat{BIAS}(\hat{\beta}) = 2\hat{\beta} - \frac{\sum_{m=1}^{M} \hat{\beta}^{*m}}{M}$$

III. SIMULATION STUDY

Our main objective is to study the behavior of the proposed bias-corrected estimator for the regression coefficients. In order to do so, we have generated the values of the variable of interest T with the model

$$\ln T = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \qquad (4)$$

where X_1 and X_2 are uniform $(0, 5)$ random, variables, $\beta_0 = \beta_1 = \beta_2 = 1$, and the ϵ's are assumed to be normally distributed with mean 0 and standard deviation taking different values $\sigma = \{1, 0.75, 0.5\}$. To be able to consider three

TABLE III
ESTIMATED BIASES (BIAS) AND VARIANCES (VAR) FOR THE ESTIMATED
COEFFICIENTS WITHOUT BIAS CORRECTION ($\hat{\beta}$), VERSUS THE
BIAS-CORRECTED PROPOSAL ESTIMATOR ($\hat{\beta}_c$) FOR DIFFERENT VALUES OF
$\sigma = \{1, 0.75, 0.5\}$ AND 15% CENSORING LEVEL.

σ		β_0 BIAS	β_0 VAR	β_1 BIAS	β_1 VAR	β_2 BIAS	β_2 VAR
1	$\hat{\beta}$	0.1849	0.197	-0.0545	0.016	-0.0636	0.014
	$\hat{\beta}_c$	-0.0292	0.208	-0.0096	0.017	-0.0107	0.016
0.75	$\hat{\beta}$	0.1009	0.110	-0.0281	0.009	-0.0344	0.008
	$\hat{\beta}_c$	-0.0169	0.113	-0.0034	0.009	-0.0047	0.009
0.5	$\hat{\beta}$	0.0430	0.049	-0.0123	0.004	-0.0139	0.003
	$\hat{\beta}_c$	-0.0117	0.051	-0.0008	0.004	-0.0003	0.004

censoring levels (i.e., 15%, 30% and 50%), we use different uniform distributions. The results have been obtained for a sample of size $n = 40$ and they are based on 1000 simulated data sets. In each data set, we have used $M = 199$ bootstrap replicates. Efron and Tibshirani [17] indicate that 200 bootstrap replications are enough for estimating the standard error and bias. The estimated biases (BIAS) and variances (VAR) for the estimated coefficients without bias correction ($\hat{\beta}$), versus the bias-corrected proposal estimator ($\hat{\beta}_c$), for different values of $\sigma = \{1, 0.75, 0.5\}$ are presented, for different censoring levels, in Tables 1 to 3.

From Tables 1 to 3, we can see that the proposed bias-corrected estimator shows a smaller bias than the one without bias correction in all considered situations. In addition, if we compute the univariate mean squared error corresponding to each estimated β_j, we can see that mean squared errors are smaller for the bias-corrected proposal in all considered cases. Therefore, if we analyze the global estimation performance, using as an indicator the multivariate or total mean square error, the same results is obtained. As can be seen the advantage of using the proposed bias-corrected estimator is more evident when the censoring level increases and/or the value of σ is larger. In addition, for the estimators considered in the simulations, the bias and variance for each coefficient decrease when the value for the parameter σ decreases, which also decreases the univariate and multivariate mean squared errors. Finally and as expected, the effect of the censoring percentage, tends to increase the variance and the bias of the estimations.

IV. CONCLUSIONS AND FINAL RECOMMENDATIONS

In this paper, we propose a bias improvement for estimating the regression parameters in a linear regression model where the response variable is censored. This proposed improvement consist on reducing the estimators' bias when there is censoring. The main motivation to used the proposed methodology in Stute [9] for censored regression models, lies on the fact that he proposes the use of a flexible model, without assuming any probability distribution, and without assuming proportional hazard functions, which, sometimes, could be a very restrictive assumption. Simulation results presented in the paper show that the proposed new estimator reduces the bias and the mean squared error of the estimation and the advantage of using

it is greater for cases with large censoring levels, where, as expected, the problem with the bias is more evident.

REFERENCES

[1] D.R. Cox, *Regression models and life-tables*, J. R. Stat. Soc. Ser. B. 34, 1972, pp. 187-220.

[2] J.F. Lawless, *Statistical Models and Methods for Lifetime Data*, John Wiley and Sons, New York, 1982.

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

[4] A.A. Tsiatis, *Estimating regression parameters using linear rank tests for censored data*, Ann. Statist. 18, 1990, pp. 354-372.

[5] T.L. Lai, and Z. Ying, *Linear rank statistics in regression analysis with censored or truncated data*, J. Multivariate Anal. 40, 1992, pp. 13-45.

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

[7] R.G. Miller, *Least squares regression with censored data*, Biometrika 63, 1976, pp. 449-464.

[8] J.J. Buckley, and I.R. James, *Linear regression with censored data*, Biometrika 66, 1979, pp. 429-436.

[9] W. Stute, *Consistent estimation under random censorship when covariables are present,* J. Multivariate Anal. 45, 1993, pp. 89-103.

[10] J. Orbe, E. Ferreira, and V. Núñez-Antón, *Censored partial regression*, Biostatistics 4, 2003, pp. 109-121.

[11] W. Stute, *Nonlinear censored regression*, Statist. Sinica 9, 1999, pp. 1089-1102.

[12] E.L. Kaplan, and P. Meier, *Nonparametric estimation from incomplete observations*, J. Amer. Statist. Assoc. 53, 1958, pp. 457-481.

[13] W. Stute, *Distributional convergence under random censorship when covariables are present*, Scand. J. Stat. 23, 1996, pp. 461-471.

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

[15] D. Mauro, *A combinatoric approach to the Kaplan-Meier estimator*, Ann. Statist. 13, 1985, pp. 142-149.

[16] W. Stute, *The bias of Kaplan-Meier integrals*, Scand. J. Stat. 21, 1994, pp. 475-484.

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