Revista Colombiana de Estadística
January 2017, Volume 40, Issue 1, pp. 85 to 103
DOI: http://dx.doi.org/10.15446/rce.v40n1.55807

Confidence Bands for the Survival Function Using a Weibull Regression Model in Presence of Arbitrary Censoring

Bandas de confianza para la función de supervivencia usando la ONU modelo de regresión de Weibull en presencia de censura arbitraria

Mario César Jaramillo Elorza[1,a], Juan Carlos Salazar Uribe[1,b]

[1]Statistics School, Universidad Nacional de Colombia, Medellín, Colombia

Abstract

Usually, the exact time at which an event occurs cannot be observed for several reasons; for instance, it is not possible to constantly monitor a characteristic of interest. This generates a phenomenon known as censoring that can be classified as having a left censor, right censor or interval censor. When one is working with survival data in the presence of arbitrary censoring, the survival time of interest is defined as the elapsed time between an initial event and the next event that is generally unknown. This problem has been widely studied in the statistic literature and some progress has been made, toward resolving and the formulation of a bivariate likelihood to estimate parameters in a parametric regression model offers positive development opportunities. In this paper, we construct a bivariate likelihood for the Weibull regression model in the presence of interval censoring. Finally, its performance is illustrated by means of a simulation study.

***Key words*:** Biostatistics, Confidence Bands, Goodness of Fit, Regression Models, Simulation, Survival Analysis.

Resumen

Usualmente, el tiempo exacto en el que ocurre un evento no se puede observar por diversas razones; por ejemplo, no es posible un monitoreo constante de las características de interés. Esto genera un fenómeno conocido como censura que puede ser de tres tipos: a izquierda, a derecha, o de intervalo. En datos de tiempo de vida con censura arbitraria (censura a izquierda, a derecha, o de intervalo), el tiempo de supervivencia de interés es definido como el lapso de tiempo entre un evento inicial y el evento siguiente, el cuál

[a]PhD. E-mail: mcjarami@unal.edu.co
[b]PhD. E-mail: jcsalaza@unal.edu.co

generalmente es desconocido. Este problema ha sido ampliamente estudiado en la literatura estadística, y se evidencian avances importantes. Sin embargo, la construcción de una verosimilitud bivariada para la estimación de los parámetros de modelos de regresión paramétricos, ofrece oportunidades de desarrollo. En este trabajo se construye una verosimilitud bivariada para el modelo de regresión Weibull, en presencia de censura arbitraria. Finalmente se ilustra su desempeño por medio de un estudio de simulación.

***Palabras clave*:** análisis de supervivencia, bandas de confianza, bioestadística, modelos de regresión, simulación.

1. Introduction

Situations in which the observed response for each individual under study is either an exact survival time or a censoring time are commonly in practice, There are two types of censoring known as type I and type II, respectively. Type I censoring occurs when units that have not failed are removed from a test at a prespecified time; on the other hand, type II censoring occurs when a life test is terminated after r failures are observed. Nonetheless, there can be situations, such as in longitudinal studies which individuals are monitored during a fixed period of time or periodically visited during a certain period. In this context, the time T_i, $i = 1, \ldots, n$, until the occurrence of an event of interest for each individual is unknown, the only known fact is that the event occurred on an interval between visits, in other words, between the visit at time L_i and the visit at time U_i, where $L_i < U_i$. It is important to point out that in such studies, the survival time T_i is not exactly known. We only know that the event of interest occurred inside the interval $(L_i, U_i]$ with $L_i < T_i \leq U_i$. Moreover, taking into account that if the event occurs in the exact moment of a visit, which is very unlikely but could occur, an exact survival time would be obtained. In this case, it is possible to assume that $L_i = T_i = U_i$.

On the other hand, it is known that for individuals whose times are right-censored, the event of interest has not occurred until the last visit, but that could happen at any time beyond that moment. Therefore, it is assumed in this case that T_i could occur inside the interval (L_i, ∞), with L_i being equal to the time period from the start of the study to the last visit, so $U_i = \infty$.

Similarly, it is known that for individuals whose times are left-censored, the event of interest occurred before the first visit and, hence, it is possible to assume that T_i occurred on the interval $(0, U_i]$ with $L_i = 0$ representing the start of the study and U_i being the time period from the start of the study to the first visit.

In survival data analysis, it is of interest to estimate the survival function $S(t)$ and to assess the importance of the potential factors or individual features over this survival time.

A common practice among data analysts is to assume that the event that already occurred inside the interval $(L_i, U_i]$ occurred either in the inferior or superior limit or in the midpoint of each interval. Some authors, such as Rucker & Messerer (1988), Odell, Anderson & D'Agostinho (1992), and Dorey, Little &

Schenker (1993), state that to treat the survival time of interest as if it were exact could lead to biased estimators as well as to partial and unreliable conclusions and estimations.

These affirmations somehow motivate different proposals related to the treatment of these censorings in order to avoid bias and to extract more information from the data. Our proposal partially covers this objective.

For the case of right censoring, the Kaplan-Meier estimator could be used to obtain $F(t)$ Kaplan & Meier (1958). However, with interval-censored data, the classic Kaplan-Meier method could not be implemented. For these interval-censored data, Peto (1973), Turnbull (1974) and Turnbull (1976) developed the so called nonparametric maximum likelihood estimator (NPMLE); from now on, we will call it Turnbull estimator.

The Turnbull estimator is based on a sample of observed intervals $[L_i, R_i]$, $i = 1, 2, \ldots n$, including the independent random variables $T_1, T_2, \ldots, T_n$. As stated above, an exact observation of T_i is obtained only if $L_i = R_i$.

Given this example, the likelihood function to be maximized is:

$$L(F) = \prod_{i=1}^{n} [F(R_i+) - F(L_i-)] \tag{1}$$

To solve this maximization problem, (Peto 1973) defines two sets: $\gamma = \{L_i, i = 1, 2, \ldots n\}$ and $\kappa = \{R_i,\ i = 1, 2, \ldots, n\}$ containing the right and left sides of the intervals respectively.

From these sets, new $[q_1, p_1]$, $[q_2, p_2]$, $\ldots$, $[q_m, p_m]$ disjoint intervals are formed, such that $q_j \in \gamma$, $p_j \in \kappa$ *and* $q_j \leq p_j$. It could be proved that a function that maximizes (1) is constant between the intervals $[q_j, p_j]$ and it is not defined inside those intervals. This implies that $\hat{P}(T \in (p_{j-1}, q_j)) = 0$ for any j. Let s_j be the increases of F inside the $[q_j, p_j]$ intervals, $j = 1, \ldots, m$, $L(F)$ must be maximized as a function of $s_1, s_2, \ldots, s_m$ subject to the constrain $s_j \geq 0$ and $s_m = 1 - \sum_{j=1}^{m-1} s_j$. Peto addresses this maximization problem using the Newton-Raphson algorithm.

In contrast to Peto, Turnbull (1976), proposes the use of the so called self-consistency algorithm for the same maximization problem. The idea of the self-consistency algorithm was first introduced by Efron (1967). Its application for the maximization in (1) is as follows: Let $\alpha_{ij} = I_{\{[q_j, p_j] \in [L_i, R_i]\}}$, $i = 1, \ldots, n$, $j = 1, \ldots m$, be the indicator variables that tell whether the interval $[q_j, p_j]$ is contained in the interval $[L_i, R_i]$. Then, the probability of T_i being inside the interval $[q_j, p_j]$, given a vector $\boldsymbol{s} = (s_1, s_2, \ldots, s_m)'$, is given by:

$$\mu_{ij}(\boldsymbol{s}) = \frac{\alpha_{ij} s_j}{\sum_{k=1}^{m} \alpha_{ik} s_k} \tag{2}$$

Since $\hat{F}$ is constant outside the intervals $[q_j, p_j]$, the proportion of observations inside the interval $[q_j, p_j]$ is equal to:

$$\pi_j(\boldsymbol{s}) = \frac{1}{n}\sum_{i=1}^{n}\mu_{ij}(\boldsymbol{s}) \tag{3}$$

And a vector $\boldsymbol{s} = (s_1, s_2, \ldots, s_m)'$ is called self-consistent if,

$$s_j = \pi_j(\boldsymbol{s}), \quad j = 1, 2, \ldots, m$$

Following this definition, the Turnbull self-consistency algorithm to calculate the nonparametric estimator of $F(t)$ could be implemented following these steps:

1. Obtain initial estimations of $\boldsymbol{s}$; for instance, $s_j^{(0)} = \frac{1}{m}$, $j = 1, 2, \ldots, m$.
2. For $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, calculate $\mu_{ij}\left(\boldsymbol{s}^{(0)}\right)$ in accordance with (2), then update $\pi_j\left(\boldsymbol{s}^{(0)}\right)$ in accordance with (3).
3. Obtain improved estimations for $\boldsymbol{s}$ by finding $s_j^{(1)} = \pi_j\left(\boldsymbol{s}^{(0)}\right)$.
4. Return to step 2, replace $\boldsymbol{s}^{(0)}$ with $\boldsymbol{s}^{(1)}$, and continue until convergence is attained.

Meeker & Escobar (1992) proposed evaluating the effect of perturbations on the model, or the weight they have on the maximum likelihood estimates obtained from censored survival data. Waller & Turnbull (1992) analyzed several graphic methods to check goodness of fit in the case of right censored data, and the proposed making an empirical rescaling of the axes to prevent data to be grouped around particular areas in the graphics. Chang & Weissfeld (1999) proposed two diagnostic methods to evaluate the accuracy of the confidence region based on the partial likelihood function using a Cox's proportional hazards model with censored data. Joly & Commenges (1999) studied both the intensity and survival function for a progressive right-shift multi-state model using arbitrary censored data; they illustrated their method using longitudinal data about AIDS. Rosales & Salazar (2006) generalized the model proposed by Joly & Commenges (1999) and formulated a likelihood function that considers the presence of arbitrary censoring. However, the problem of constructing simultaneous confidence bands with arbitrary censoring still presents opportunities for development. This paper discusses how to obtain simultaneous confidence bands when a Weibull regression model with arbitrary censoring is considered. In the case of simultaneous confidence bands (SCB) for the cumulative distribution function, Cheng & Iles (1983) used the Wald statistic to construct the SCB for quantiles of the cumulative distribution function and the probability of failure. Cheng & Iles (1988) extended their confidence bands results to cumulative distribution functions that are members of the location and scale family with complete data. Jeng & Meeker (2001) generalized the work of Cheng & Iles (1988) using the Wald statistic with the

observed Fisher's information matrix, the Wald statistic with local information matrix Fisher, and the likelihood ratio statistic. Finally, Escobar, Hong & Meeker (2009) extend the work of Cheng & Iles (1983) in the following ways:

1. They showed how to compute SCB based on local information, expected information, and estimated expected information for both the "cdf method" and the "quantile method", Escobar et al. (2009); Cheng & Iles (1983) considered only the expected information case

2. They described a calibration method of the intervals to provide exact coverage for type II censoring and improved approximate coverage for other kinds of censoring.

3. They discussed how to extend these procedures to regression analysis.

This work was motivated by a radiographic progression study conducted in Colombia the propose of which was to identify risk factors related with Rheumatoid Arthritis (RA Rojas, Diaz, Calvo, Salazar, Iglesias, Mantilla & Anaya 2009). Suppose that a patient is observed at irregular times and at each visit his/her health state is assessed and classified in three categories namely mild, moderate, or severe. Since, in general, it is not possible to observe a patient continuously, one of the following situations would probably by time:

1. On the first visit, the patient could be in a moderate or severe state of the disease. In this case, the time when the patient changed from mild to moderate or from mild to severe is unknown. This generates left censored data.

2. The patient is observed at least once in mild or moderate condition and then he/she left the study for some reason. This generates right censored data.

3. In two consecutive visits the patient changed of state (say from mild to moderate or from moderate to severe) but the exact time when this occurred is unknown. This generates interval censored data.

This dataset about radiographic progression of RA exhibits these three types of censoring, and therefore it is not convenient to analyze it using conventional approaches that take into account only right censored data, such as the well-known Cox model. Even if we fit a parametric model that takes into account the dynamics of censoring the data set, the way the goodness of fit is evaluated could not be entirely correct because the confidence bands of Nair (1984) are used; these are nonparametric and only work for right censored data. Ii then seems more reasonable to build confidence bands that take into account arbitrary censored data. PROC LIFEREG of SAS©, allows data to be modeled with censored arbitrary data as long as a parametric regression model is specified. Allison (1995) fitted a Weibull model, but the way he assessed the model's goodness of fit is not entirely satisfactory because he used Nair's confidence bands in the presence of interval censored data, which cannot not correct.

The goal of this paper is to propose simultaneous confidence bands for a Weibull regression model in the presence of arbitrary censored data. Specifically, instead of using likelihood to obtain the confidence interval, we adapted the simultaneous confidence parametric bands proposed by Escobar et al. (2009) in conjunction with the likelihood function of a bivariate distribution. This is a different strategy from that of just imputing the interval censored data. This strategy is the works most important contribution and it yields simultaneous parametric confidence bands. It is contrasted this with PROC LIFEREG of SAS©, which yields nonparametric confidence bands.

They also made comparisons, using a simulation study based on the deviance of two models. The first estimated the parameters using likelihood with arbitrary censoring, and the other estimated the parameters using a bivariate likelihood (Gentleman & Vandal 2001). The goal was to evaluate which of the two likelihoods yielded better estimates.

Take into consideration that the goal of this paper is to build a bivariate likelihood with dependency for interval-censored data in order to find $\widehat{S}(t)$. Since this dependency will be specified by means of copulas, it is important to first define them.

2. Copulas

Suppose that C_α is a distribution function with density c_α over $[0,1]^2$ for $\alpha \in \mathbb{R}$. Let (T_1, T_2) be the failure times and let both (S_1, S_2) and (f_1, f_2) be its corresponding marginal survival and density functions, respectively. If (T_1, T_2) comes from a copula C_α, for any $\alpha \in \mathbb{R}$, the joint survival and density functions of (T_1, T_2) are given by

$$S(t_1, t_2) = C_\alpha(S_1(t_1), S_2(t_2)) \quad t_1, t_2 \geq 0,$$

$$f(t_1, t_2) = c_\alpha(S_1(t_1), S_2(t_2)) f_1(t_1) f_2(t_2) \quad t_1, t_2 \geq 0,$$

where α represents the dependency parameter between T_1 and T_2.

We will use the Archimedean family of copulas because is the most used copula family. A bivariate distribution belonging to the family of Archimedean copula models can be represented in the following way:

$$C_\alpha(u, v) = \phi_\alpha^{-1}[\phi_\alpha(u) + \phi_\alpha(v)], \quad 0 \leq u, v \leq 1,$$

where ϕ is a convex and decreasing function such that $\phi \geq 0$, $\phi(1) = 0$. The ϕ function is named *generator* of the C_α copula and the inverse of the generator, ϕ^{-1} and is the *Laplace transform* of a latent variable denoted as γ, which induces the dependency α. Thus, the selection of a generator results in several families of copulas. Table 1 shows the forms for bivariate survival functions in three Archimedean copula families. Additionally, Table 2 shows the generators and the Laplace transform for the considered families.

Table 1: Common Archimedean copulas.

Family Copula	Parameter Space	Bivariate Copula $C_\alpha(u,v)$
Clayton	$\alpha > 1$	$\left\{u^{1-\alpha} + v^{1-\alpha} - 1\right\}^{1/(1-\alpha)}$
Gumbel	$0 < \alpha < 1$	$\exp\left\{-\left[(-\ln u)^{1/\alpha} + (-\ln v)^{1/\alpha}\right]^{\alpha}\right\}$
Frank	$\alpha > 0$	$\log_\alpha \left\{1 + (\alpha^u - 1)(\alpha^v - 1)/(\alpha - 1)\right\}$

Table 2: Generators and their Laplace Transforms.

Family Copula	Parameter Space	Generator $\phi(t)$	Laplace Transform $(\tau(s) = \phi^{-1}(s))$
Clayton	$\alpha > 1$	$t^{1-\alpha} - 1$	$(1+s)^{1/(1-\alpha)}$
Gumbel	$0 < \alpha < 1$	$(-\ln t)^{1/\alpha}$	$\exp(-s^\alpha)$
Frank	$\alpha > 0$	$\ln \frac{\alpha^t - 1}{\alpha - 1}$	$\log_\alpha \{1 - (1-\alpha)e^s\}$

3. A Likelihood Function for Interval-Censored Bivariate Data

Let T and V be two random variables with the cumulative distribution function $F(t, v)$; both T and V are Type I interval censoring. So, instead of observing the pair (T, V), we observe the vector $\mathbf{\Psi} = (T_1, T_2, V_1, V_2, \mathbf{\Delta})$. Here $0 < T_1 < T_2 < \infty$ are the observation times for T, $0 < V_1 < V_2 < \infty$ are the observation times for V, and $\mathbf{\Delta}$ is the vector $\mathbf{\Delta} = (\Delta_{11}, \Delta_{12}, \Delta_{13}, \Delta_{21}, \Delta_{22}, \Delta_{23}, \Delta_{31}, \Delta_{32}, \Delta_{33})$. Δ_{jk} is defined as:

$\Delta_{11} = I_{\{T \le T_1, V \le V_1\}}$
$\Delta_{12} = I_{\{T_1 < T \le T_2, V \le V_1\}}$
$\Delta_{13} = I_{\{T > T_2, V \le V_1\}}$
$\Delta_{21} = I_{\{T \le T_1, V_1 < V \le V_2\}}$
$\Delta_{22} = I_{\{T_1 < T \le T_2, V_1 < V \le V_2\}}$
$\Delta_{23} = I_{\{T > T_2, V_1 < V \le V_2\}}$
$\Delta_{31} = I_{\{T \le T_1, V > V_2\}}$
$\Delta_{32} = I_{\{T_1 < T \le T_2, V > V_2\}}$
$\Delta_{33} = I_{\{T > T_2, V > V_2\}}$

Let $\boldsymbol{T} = (T_1, T_2)$ and $\boldsymbol{V} = (V_1, V_2)$ be two bivariate random variables with the joint probability density function $g(\mathbf{t}, \mathbf{v})$, the joint cumulative distribution function $G(\mathbf{t}, \mathbf{v})$. Also $\boldsymbol{t} = (t_1, t_2)$ y $\boldsymbol{v} = (v_1, v_2)$ are the respective observations of these variables.

$R_{ij}(\boldsymbol{t}, \boldsymbol{v})$ is defined as a function from $\mathbb{R}^4_+$ to $\mathbb{R}^2_+$ where $\boldsymbol{t} = (t_1, t_2)$ and $\boldsymbol{v} = (v_1, v_2)$, as follows:

$R_{11}(\boldsymbol{t}, \boldsymbol{v}) = [0, t_1] \times [0, v_1]$
$R_{12}(\boldsymbol{t}, \boldsymbol{v}) = (t_1, t_2] \times [0, v_1]$
$R_{13}(\boldsymbol{t}, \boldsymbol{v}) = (t_2, \infty) \times [0, v_1]$
$R_{21}(\boldsymbol{t}, \boldsymbol{v}) = [0, t_1] \times (v_1, v_2]$
$R_{22}(\boldsymbol{t}, \boldsymbol{v}) = (t_1, t_2] \times (v_1, v_2]$

$R_{23}(\boldsymbol{t},\boldsymbol{v}) = (t_2,\infty)\times(v_1,v_2]$
$R_{31}(\boldsymbol{t},\boldsymbol{v}) = [0,t_1]\times(v_2,\infty)$
$R_{32}(\boldsymbol{t},\boldsymbol{v}) = (t_1,t_2]\times(v_2,\infty)$
$R_{33}(\boldsymbol{t},\boldsymbol{v}) = (t_2,\infty)\times(v_2,\infty)$

It is understood that $(\boldsymbol{T},\boldsymbol{V})$ and (T,V) are independent and $\Pr(T_1<T_2) = \Pr(V_1<V_2)=1$.

We assume n independent and identically distributed repetitions of $\boldsymbol{\Psi}$. Notice that $\Pr(U_1<U_2)=\Pr(V_1<V_2)=1$. The underlying repetitions of (T,V) are $(t_1,v_1),\ldots,(t_n,v_n)$. For each observation i, the $(\boldsymbol{T}_i,\boldsymbol{V}_i)$ points define 9 rectangles the are denoted as R_{jki}, for $j,k=1,2,3$, where the values of $\boldsymbol{\Delta}_i = (\Delta_{11i},\Delta_{12i},\Delta_{13i},\Delta_{21i},\Delta_{22i},\Delta_{23i},\Delta_{31i},\Delta_{32i},\Delta_{33i})$ indicate which rectangle includes the (t_i,v_i) pair.

Let $g(\boldsymbol{t},\boldsymbol{v})$ denote the joint density of $(\boldsymbol{T},\boldsymbol{V})$, where $\boldsymbol{t}=(t_1,t_2)$ and $\boldsymbol{v}=(v_1,v_2)$. Let $f(t,v)$ denote the joint density of (T,V). Since $(\boldsymbol{T},\boldsymbol{V})$ and (T,V) are independent, the joint density of $(\boldsymbol{T},\boldsymbol{V},T,V)$ is $h(\boldsymbol{t},\boldsymbol{v},t,v)=g(\boldsymbol{t},\boldsymbol{v})f(t,v)$. Thus, using the notation $R(\boldsymbol{t},\boldsymbol{v})=R(t_1,t_2,v_1,v_2)$ and the fact that $\Delta_{11}=1$, the distribution of $\boldsymbol{\Psi}$ is obtained as follows:

$$
\begin{aligned}
F_{\boldsymbol{\Psi}}(\boldsymbol{\psi}) &= \Pr(T_1\le t_1, T_2\le t_2, V_1\le v_1, V_2\le v_2, \Delta_{11}=1)\\
&= \Pr(T_1\le t_1, T_2\le t_2, V_1\le v_1, V_2\le v_2, T\le T_1, V\le V_1)\\
&= \int_0^{v_2}\int_0^{v_1}\int_0^{t_2}\int_0^{t_1}\left[\iint_{R(\boldsymbol{t}',\boldsymbol{v}')} h(t_1',t_2',v_1',v_2',t,v)dtdv\right]dt_1'dt_2'dv_1'dv_2'\\
&= \int_0^{v_2}\int_0^{v_1}\int_0^{t_2}\int_0^{t_1} g(\boldsymbol{t}',\boldsymbol{v}')\left[\iint_{R(\boldsymbol{t}',\boldsymbol{v}')} f(t,v)dtdv\right]dt_1'dt_2'dv_1'dv_2'\\
&= \int_0^{v_2}\int_0^{v_1}\int_0^{t_2}\int_0^{t_1} g(\boldsymbol{t}',\boldsymbol{v}')\ \Pr\big[(T,V)\in R(\boldsymbol{t}',\boldsymbol{v}')\big]\,dt_1'dt_2'dv_1'dv_2'\\
&= \int_0^{v_2}\int_0^{v_1}\int_0^{t_2}\int_0^{t_1} g(\boldsymbol{t}',\boldsymbol{v}')\ \Pr_F\big[R(\boldsymbol{t}',\boldsymbol{v}')\big]\,dt_1'dt_2'dv_1'dv_2'\\
&= \int_0^{v_2}\int_0^{v_1}\int_0^{t_2}\int_0^{t_1} g(\boldsymbol{t}',\boldsymbol{v}')\ \Pr_F\big[R(\boldsymbol{t}',\boldsymbol{v}')\big]\,d\boldsymbol{t}'d\boldsymbol{v}'
\end{aligned}
$$

Here, for convenience, we use the notations $d\boldsymbol{t}'=dt_1'dt_2'$ and $d\boldsymbol{v}'=dv_1'dv_2'$.

It can be concluded that the density of $\boldsymbol{\Psi}$ is $g(\boldsymbol{t},\boldsymbol{v})\ \Pr_F\big[R(\boldsymbol{t},\boldsymbol{v})\big]$, where $g(\boldsymbol{t},\boldsymbol{v})$ is independent of F.

Usually, if $\Delta_{jk}=1, (j,k)\in\{1,2,3\}^2$, the density of $\boldsymbol{\Psi}$ is: $g(\boldsymbol{t},\boldsymbol{v})\ \Pr_F\big[R(\boldsymbol{t},\boldsymbol{v})\big]$, where $g(\boldsymbol{t},\boldsymbol{v})$ is independent of F. Then, the likelihood function of F is: $L_n(F) = \prod_{i=1}^n\prod_{j,k=1}^3\big\{\Pr_F\big[R_{jk}(\boldsymbol{t},\boldsymbol{v})\big]\big\}^{\delta_{jki}}$.

Then, the loglikelihood is: $\ell_n(F)=\sum_{i=1}^n\sum_{j,k=1}^3\delta_{jki}\log\{\Pr_F\big[R_{jk}(\boldsymbol{t},\boldsymbol{v})\big]\}$. If F_T is the marginal distribution function for T and F_V is the marginal distribution function for V, the loglikelihood for F is given by:

$$\begin{aligned}\ell_n(F) = \sum_{i=1}^{n} \{ & \delta_{11i} \log[F(t_{1i,}, v_{1i})] + \delta_{12i} \log[F(t_{2i}, v_{1i}) - F(t_{1i}, v_{1i})] + \delta_{13i} \log[F_2(v_{1i}) \\ & - F(t_{2i}, v_{1i})] + \delta_{21i} \log[F(t_{1i}, v_{2i}) - F(t_{1i}, v_{1i})] + \delta_{22i} \log[F(t_{2i}, v_{2i}) \\ & - F(t_{1i}, v_{2i}) - F(t_{2i}, v_{1i}) + F(t_{1i}, v_{1i})] + \delta_{23i} \log[F_2(v_{2i}) - F(t_{2i}, v_{2i}) \\ & - F_2(v_{1i}) + F(t_{2i}, v_{1i})] + \delta_{31i} \log[F_1(t_{1i}) - F(t_{1i}, v_{2i})] \\ & + \delta_{32i} \log[F_1(t_{2i}) - F_1(t_{1i}) - F(t_{2i}, v_{2i}) \\ & + F(t_{1i}, v_{2i})] + \delta_{33i} \log[1 - F_1(t_{2i}) - F_2(v_{2i}) + F(t_{2i}, v_{2i})]\}\end{aligned}$$

When we only have interval and right censoring, $\delta_{11i} = 0$, $\delta_{12i} = 0$ and $\delta_{13i} = 0$, then $\ell_n(F)$ reduces to:

$$\begin{aligned}\ell_n(F) = \sum_{i=1}^{n} \{ & \delta_{22i} \log[F(t_{2i}, v_{2i}) - F(t_{1i}, v_{2i}) - F(t_{2i}, v_{1i}) + F(t_{1i}, v_{1i})] + \\ & \delta_{33i} \log[1 - F_1(t_{2i}) - F_2(v_{2i}) + F(t_{2i}, v_{2i})]\}\end{aligned}$$

In terms of the survival function, we have:

$$\begin{aligned}\ell_n(S) = \sum_{i=1}^{n} \{ & \delta_{22i} \log[S(t_{1i}, v_{1i}) - S(t_{1i}, v_{2i}) - S(t_{2i}, v_{1i}) + S(t_{2i}, v_{2i})] + \\ & \delta_{33i} \log[S(t_{2i}, v_{2i})]\}\end{aligned}$$

since, $F(t, v) = 1 - S_1(t) - S_2(v) + S(t, v)$

Consider the Weibull regression model,

$$\log(T) = \beta_0 + \boldsymbol{\beta}' \boldsymbol{Z} + \sigma W$$

Where the response variable T could include the three types of censoring (left, right, and interval censoring), $\boldsymbol{\beta}$ is a vector of unknown parameters, and σ is the scale parameter ($\sigma > 0$), $T \sim$ Weibull(μ, σ), $W \sim$ SEV$(0, 1)$, with $\mu = \beta_0 + \boldsymbol{\beta}' \boldsymbol{Z}$ where SEV (0.1) is the standard smallest extreme value distribution.

To verify the assumptions of the Weibull regression model, we use the standardized residuals:

$$W_j = \frac{\log T_j - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}' \boldsymbol{Z}_j}{\hat{\sigma}}$$

If the Weibull model is suitable, then these residuals could be thought of as a censored sample of a small extreme value distribution, ($W \sim$ SEV(0,1)).

Let V be an auxiliary variable so that T and V are highly dependent, let $\tau_{T,V}$, be the Kendall's tau, τ, between T and V. Since $W = (\log T - \beta_0 - \boldsymbol{\beta}' \boldsymbol{Z})/\sigma$, is an increasing function of T and τ is invariant under monotonic transformations, $\tau_{T,V} = \tau_{W,V}$.

To estimate the parameters of the Weibull regression model, we use the bivariate loglikelihood for S, which is expressed as:

$$\ell_n(S) = \sum_{i=1}^{n} \{\delta_{22i} \log[S(w_{1i}, v_{1i}) - S(w_{1i}, v_{2i}) - S(w_{2i}, v_{1i}) + S(w_{2i}, v_{2i})] + \delta_{33i} \log[S(w_{2i}, v_{2i})]\}$$

If we assume that $V \sim \text{UNIF}(a, b)$, then

$$S_1(w) = \exp\{-\exp(w)\}, \; S_2(v) = \frac{b - v}{b - a}$$

If the Gumbel copula is used for constructing the bivariate distribution with dependency parameter τ, we have the following:

$$S(w, v) = \exp\left\{-\left[(\exp w)^{1/\alpha} + \left[-\log\left(\frac{b - v}{b - a}\right)\right]^{1/\alpha}\right]^{\alpha}\right\}$$

Even though the uniform distribution has rough edges, it works well in the simulation process, as we will show in the next section; however, another distribution could be used, for instance, the beta distribution.

4. Simulation Study

To explore if the bivariate likelihood with random censoring improves the estimations of the parameters of the Weibull regression model in comparison to the ones obtained with the Turnbull method (Turnbull 1976), the following simulation study was carried out.

Recall that the Weibull regression model is specified as:

$$\log(T) = \beta_0 + \beta Z + \sigma W$$

Therefore, to generate times from a Weibull model, we must generate Z and W, maintaining β, β_0 and σ fixed. Since it is assumed that T and V are highly dependent and that their dependence could be measured with the coefficient τ, that was fixed at $\tau = 0.99$, and since τ is invariant under monotonic transformations, then we must generate W in such a way it satisfies $\tau(W, V) = 0.99$.

The simulation factors that will be controlled are:

1. Sample size n: the objective of this factor is to assess the effect of the number of individuals in the study during the estimation process. Values of $n = 50, 100, 200$ will be taken because they can easily appear in practice.

2. Percentage of interval censored observations p: the objective of this factor is to evaluate the effect of the percentage of interval censored observations p during the estimation process. The values of $p = 0.5, 0.7, 0.9$, will reflect situations with high interval censoring percentages; the remaining data are right censored.

3. Variance of time until the event of interest σ_T^2: the objective of this factor is to assess the effect of the variance of the time until the event of interest during the estimation process. The values of $\sigma_T^2 = 4, 25, 100$ because we want to observe the effect in the presence of small and large variances.

4. Parameter vector β: the objective of this factor is to evaluate the effect of the explanatory variable Z, on the estimation process. We will consider the following values: $\beta = -0.9, -0.7, -0.5, -0.3$. Some simulations were performed with positive values of β and they yielded very similar results to the ones obtained using negative values.

5. Distribution of the explanatory variable Z: the objective of this factor is to assess the effect of the distribution of the explanatory variable Z on the estimation process. For the sake of simplicity only two distributions will be considered, a continuous standard normal distribution $Z \sim \text{NOR}(0, 1)$ and an ordinal discrete binomial distribution with parameters $n = 6$ and $p = 0.5$, ($Z \sim \text{BIN}(6, 0.5)$). However, more complex distributions could also be considered.

Finally, with the simulated data, β_0, β and σ will be estimated to obtain $\widehat{\beta_0}, \widehat{\beta}, \widehat{\sigma}$, and the square root of the mean square errors will be calculated to observe the accuracy of the estimation process.

In Lawless & Babineau (2006), we find a comprehensive discussion on how to generate interval-censored data. Ii specifically refers to the case of a longitudinal study, in which there is a periodic follow-up of the scheduled visits, and it takes into account that the patients could miss some of their appointments. We supposed that there are M potential inspection times a_j, $j = 0, \ldots, M$, for instance $a_j = j$. The probability of patients attending each scheduled visit is p. For an individual i, the observed interval censored $(L_i, R_i]$ is constructed by defining R_i as the first visit in which the event of interest is observed. L_i is the previous visit, i.e. $L_i = \max a_j : a_j < T_i, \delta_j^i = 1$ and $R_i = \min a_j : a_j \leq T_i, \delta_j^i = 1$, where $\delta_j^i = 1$, indicates that the visit occurred at time a_j. Different values of p lead to different interval lengths. For instance, $p = 0.3$ implies that 70% of the visits are missed, which would lead to the observation of wide confidence intervals for T.

With censored data, β_0, β and σ will be estimated using the interval-censored likelihood and we will write those estimates as $\widehat{\beta_{0\text{int}}}, \widehat{\beta}_{\text{int}}$ and $\widehat{\sigma}_{\text{int}}$, then the square root of the mean square errors will be calculated to measure the accuracy of the estimation process.

With the censored data, β_0, β and σ, will be estimated by taking the likelihood as a bivariate likelihood and we will denote those estimates as $\widehat{\beta_{0\text{biv}}}, \widehat{\beta}_{\text{biv}}$ and $\widehat{\sigma}_{\text{biv}}$. The square root of the mean square errors will they be calculated to measure the accuracy of the estimation process. This optimization process will be carried out using the Nelder-Mead Simplex algorithm (Nelder & Mead 1965), which is one of the options included in the `maxLik` package of R software. This algorithm was used instead of the Newton-Raphson method because it showed a better performance in the preliminary tests.

Additionally, in each simulation, deviance will be calculated using two likelihoods, one with interval censoring to estimate three parameters, β_0, β, σ, and the other, the bivariate likelihood for calculating μ, β, σ, a and b of the distribution of the V auxiliary variable, which we assumed has a UNIF(a, b) distribution. We the calculated $D = -2 \times [l(\widehat{\beta_0}_{\text{int}}, \widehat{\beta}_{\text{int}}, \widehat{\sigma}_{\text{int}}) - l(\widehat{\beta_0}_{\text{biv}}, \widehat{\beta}_{\text{biv}}, \widehat{\sigma}_{\text{biv}}, \widehat{a}, \widehat{b})]$, in this case, the approximate distribution of D is a chi-square with 2 degrees of freedom, $D \sim \chi^2_{(2)}$. The number of times H_0 is rejected at a $\alpha = 0.05$ level will also be calculated. In other words, this will be the number of times in which the bivariate likelihood is better than the interval likelihood: we will call it "acceptance".

5. Simulation Study Results

Below, we present the square root of the mean square errors of the estimations that were obtained using the methods based on the likelihoods for β_0, β, and σ. For we used some combinations of the parameters and an explanatory variable following a normal distribution ($Z \sim \text{NOR}(0, 1)$).

In Tables 3 to 6, we observe that if we take the likelihood as a bivariate likelihood for random-censored data, taking into account the auxiliary variable V, and if we estimate β_0, β, and σ, the square root of the mean square errors associated to β_0, β, and σ are much lower than if we estimate the Weibull model parameters by using the traditional likelihood with random censoring without taking into account the auxiliary variable V. Besides, we see that the square root of the mean square errors does not significantly change when the sample size n, the censoring percentage p, the variance of the time of interest σ_T^2, or the coefficient of the explanatory variable Z, β are changed. Moreover, when we compare both the likelihoods by using the likelihood-ratio test, we observe that the percentage of times the bivariate likelihood is greater than the random censoring likelihood is close to 100%.

Table 3: MSE using $Z \sim \text{NOR}(0, 1), \sigma_T = 10, p = 0.7, \beta = -0.5$.

	n		
	50	100	200
Error$(\widehat{\beta}_{\text{int}}, \beta)$	0.7501	0.6522	0.5840
Error$(\widehat{\beta}_{\text{biv}}, \beta)$	0.6215	0.6006	0.4100
Error$(\widehat{\beta}, \beta)$	0.2734	0.2021	0.1203
Error$(\widehat{\beta_0}_{\text{int}}, \beta_0)$	2.6495	2.9158	2.9741
Error$(\widehat{\beta_0}_{\text{biv}}, \beta_0)$	0.6596	0.4370	0.3494
Error$(\widehat{\beta_0}, \beta_0)$	0.6327	0.3378	0.3213
Error$(\widehat{\sigma}_{\text{int}}, \sigma)$	0.9812	0.9816	0.9876
Error$(\widehat{\sigma}_{\text{biv}}, \sigma)$	0.2715	0.2548	0.1764
Error$(\widehat{\sigma}, \sigma)$	0.0452	0.0175	0.0030
Acceptance	0.988	1.0000	1.0000

TABLE 4: MSE using $Z \sim \text{NOR}(0,1), \sigma_T = 10, n = 200, \beta = -0.5$

	p		
	0.5	0.7	0.9
$\text{Error}(\widehat{\beta}_{\text{int}}, \beta)$	0.7810	0.5840	0.5407
$\text{Error}(\widehat{\beta}_{\text{biv}}, \beta)$	0.7423	0.4100	0.4616
$\text{Error}(\widehat{\beta}, \beta)$	0.2204	0.1203	0.2445
$\text{Error}(\widehat{\beta_0}_{\text{int}}, \beta_0)$	2.3222	2.9741	2.8367
$\text{Error}(\widehat{\beta_0}_{\text{biv}}, \beta_0)$	0.9810	0.3494	0.4382
$\text{Error}(\widehat{\beta_0}, \beta_0)$	0.1185	0.3213	0.3658
$\text{Error}(\widehat{\sigma}_{\text{int}}, \sigma)$	0.9917	0.9876	0.9706
$\text{Error}(\widehat{\sigma}_{\text{biv}}, \sigma)$	0.1914	0.1764	0.2712
$\text{Error}(\widehat{\sigma}, \sigma)$	0.0638	0.0030	0.0036
Acceptance	1.0000	1.0000	1.0000

TABLE 5: MSE using $Z \sim \text{NOR}(0,1), n = 200, p = 0.7, \beta = -0.5$.

	σ_T		
	2	5	10
$\text{Error}(\widehat{\beta}_{\text{int}}, \beta)$	0.4735	0.4724	0.5840
$\text{Error}(\widehat{\beta}_{\text{biv}}, \beta)$	0.1380	0.3456	0.4100
$\text{Error}(\widehat{\beta}, \beta)$	0.0242	0.0599	0.1203
$\text{Error}(\widehat{\beta_0}_{\text{int}}, \beta_0)$	3.2262	3.1482	2.9741
$\text{Error}(\widehat{\beta_0}_{\text{biv}}, \beta_0)$	0.4535	0.3752	0.3494
$\text{Error}(\widehat{\beta_0}, \beta_0)$	0.0729	0.1801	0.3213
$\text{Error}(\widehat{\sigma}_{\text{int}}, \sigma)$	0.9029	0.9634	0.9876
$\text{Error}(\widehat{\sigma}_{\text{biv}}, \sigma)$	0.1265	0.3215	0.1764
$\text{Error}(\widehat{\sigma}, \sigma)$	0.0002	0.0017	0.0030
Acceptance	0.9995	1.0000	1.0000

TABLE 6: MSE using $n = 200, Z \sim \text{NOR}(0,1), \sigma_T = 10, p = 0.7$.

	β			
$\text{Error}(\widehat{\beta}_{\text{int}}, \beta)$	0.7717	0.5731	0.5840	0.6325
$\text{Error}(\widehat{\beta}_{\text{biv}}, \beta)$	0.6016	0.5366	0.4100	0.4112
$\text{Error}(\widehat{\beta}, \beta)$	0.1238	0.1325	0.1203	0.2221
$\text{Error}(\widehat{\beta_0}_{\text{int}}, \beta_0)$	2.9750	2.9725	2.9741	2.5948
$\text{Error}(\widehat{\beta_0}_{\text{biv}}, \beta_0)$	0.5367	0.4071	0.3494	0.4292
$\text{Error}(\widehat{\beta_0}, \beta_0)$	0.3637	0.3727	0.3213	0.3199
$\text{Error}(\widehat{\sigma}_{\text{int}}, \sigma)$	0.9829	0.9808	0.9876	0.9840
$\text{Error}(\widehat{\sigma}_{\text{biv}}, \sigma)$	0.2856	0.2930	0.1764	0.1718
$\text{Error}(\widehat{\sigma}, \sigma)$	0.0112	0.0151	0.0030	0.0110
Acceptance	1.00	1.00	1.00	0.9995

Figure 1 shows that the square root of the mean square errors does not substantially change when varying the sample size. It also shows that if we take the likelihood as a bivariate likelihood for random-censored data, taking into account the auxiliary variable V, and if we estimate β_0, β, and σ, the square root of the mean square errors of β_0, β, and σ are much lower than if we estimate these parameters of the Weibull model using a traditional likelihood with random censoring without considering the auxiliary variable V.

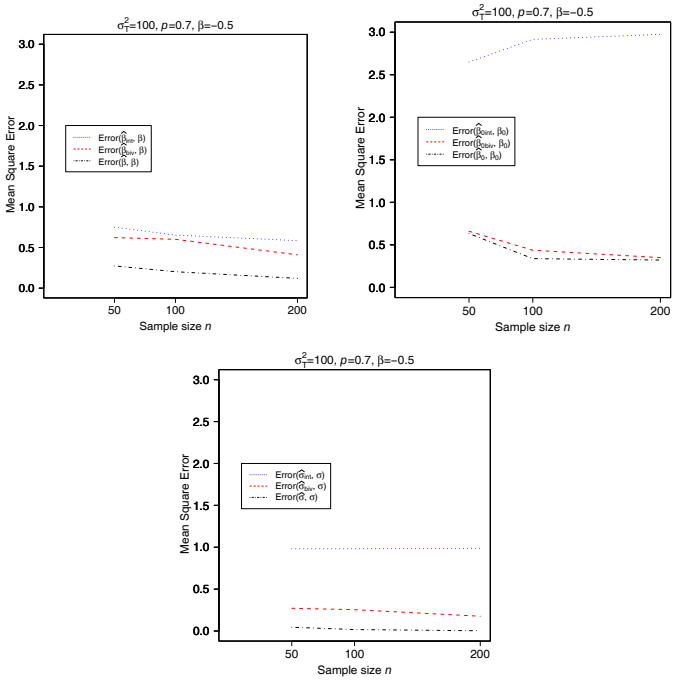

FIGURE 1: MSE behavior varying the sample size n using the three estimation methods.

Figure 2 shows that the square root of the mean square errors does not substantially change when varying the right censoring percentage p and that if we take the likelihood as a bivariate likelihood for random-censored data, taking into account the auxiliary variable V, and if we estimate β_0, β, and σ, the square root of the mean square errors of β_0, β and σ are much lower than if we estimate these parameters of the Weibull model using a traditional likelihood with random censoring without considering the auxiliary variable V.

Figure 3 shows that the square root of the mean square errors does not substantially change when varying the variance of the time of interest. It also shows that if we take the likelihood as a bivariate likelihood for random-censored data, taking into account the auxiliary variable V, and if we estimate β_0, β, and σ, the square root of the mean square errors of β_0, β, and σ are much lower than if we estimate these parameters of the Weibull model using a traditional likelihood with random censoring without considering the auxiliary variable V.

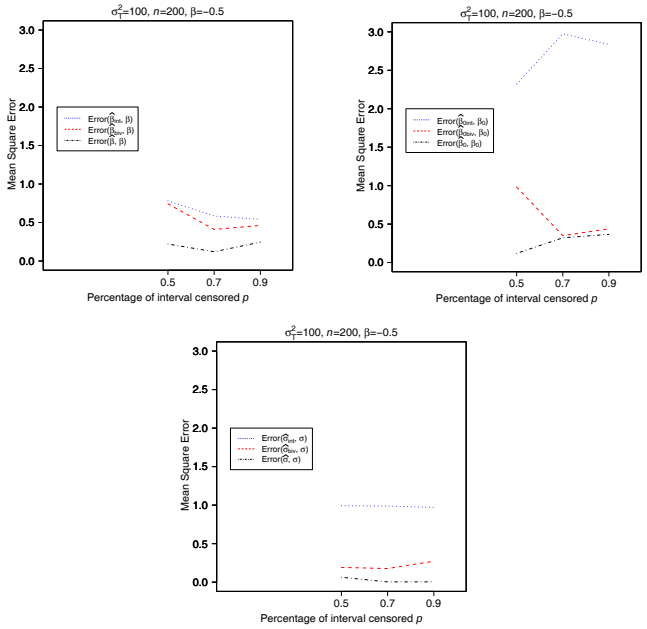

FIGURE 2: MSE behavior varying the proportion of interval censored p using three estimation methods.

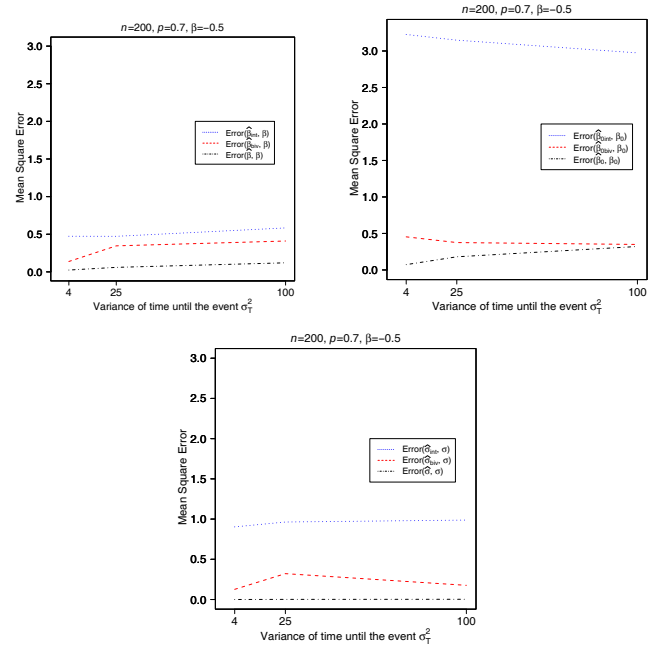

FIGURE 3: MSE behavior varying the variance of T using three estimation methods.

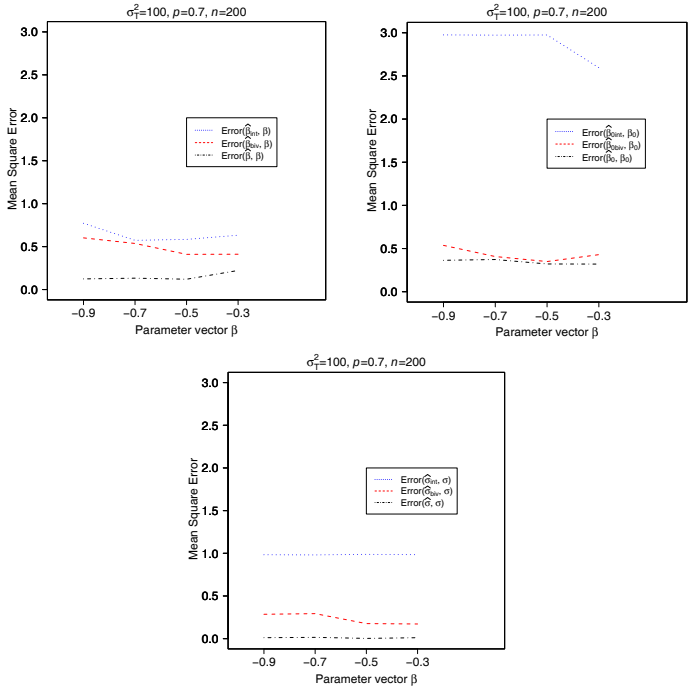

FIGURE 4: MSE behavior varying the coefficient of the explanatory variable β_0 using three estimation methods.

Figure 4 shows that the square root of the mean square errors does not substantially change when varying the coefficient of the explanatory variable β. It also shows that if we take the likelihood as a bivariate likelihood for random-censored data, taking into account the auxiliary variable V, and if we estimate β_0, β, and σ, the square root of the mean square errors of β_0, β, and σ are much lower than if we estimate these parameters of the Weibull model using a traditional likelihood with random censoring without considering the auxiliary variable V.

In Figure 5 simultaneous parametric confidence bands (Escobar et al. 2009) are shown. To construct these, we used both data with arbitrary censored, and a bivariate likelihood functions with arbitrary censored data that has an auxiliary variable V that is highly correlated with the response variable. On the right side of the graph, we can see that when the cumulative distribution function with the bivariate likelihood is estimated, taking into account the auxiliary variable V, the cumulative distribution is fairly close to the real cumulative distribution. However, if we do not take into account the auxiliary variable V, the estimated cumulative distribution is not that close to the real cumulative distribution. In the graph on the left, we can observe that the confidence parametric bands proposed by Escobar et al. (2009), in the case of the auxiliary variable, contain all the straight lines. These represent the real cumulative distribution function, whereas when if we do not take into account the auxiliary variable, this straight line is out of the

confidence bands. In summary, the use of the bivariate likelihood, the construction of which is undertaken considering the auxiliary variable, is recommended.

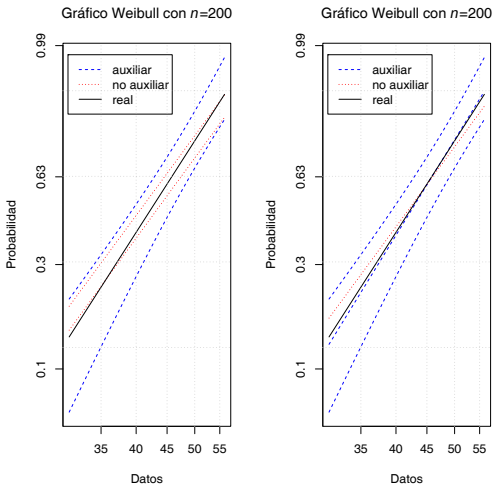

FIGURE 5: Extension of the simultaneous confidence bands of Escobar et al. (2009) for $F(t)$ to the interval censorored case, using the two likelihoods.

6. Conclusions and Recommendations

When the goal is to study the lapse of time until the occurrence of an event of interest, and to detect whether such event occurred, it is necessary to measure a variable that could be and index. This index which is known as an auxiliary variable, is correlated to the time of occurrence of the event, and this occurrence time could exhibit left, right, or interval censoring. In addition, if some covariates are available and we want to adjust a parametric regression model to determine which of these covariates are related to the time of occurrence of an event, we can not only use a likelihood with the three censoring types, but also the proposed bivariate likelihood. To obtain the maximum-likelihood estimators of $\beta_0, \boldsymbol{\beta}, \sigma$, the `maxLik` package of the R software was used as this package maximizes the likelihood functions. After this, the Nelder-Mead method was applied since it showed greater stability during the estimation process.

According to the results from the simulation study, we can conclude that the estimated parameters of the Weibull model using the proposed methodology (the bivariate likelihood) are closer to the real values of the parameters than the ones obtained only taking into consideration the three types of censoring. However, it is worth noting that the standard errors associated with the proposed method are consistently higher than the ones from the conventional methods in all the simulation scenarios.

Also, from the simulation study, we can see that, according to the likelihood ratio test, the proposed model (that uses the auxiliary variable in addition to

the three types of censoring) performs better than the model that only takes into consideration the three censoring types. This is because a higher percentage of acceptance was obtained.

According to these conclusions, when interval-censored data are available, for which the interval censoring is determined by the measurement of a variable that indicates whether the event of interest occurred or not, and when the goal is to adjust a Weibull regression model, the use of the bivariate likelihood proposed in this paper is recommended. This is because it produces closer estimations to the real parameters than the estimations obtained when the likelihood for interval-censored data is used.

In terms of future work, this methodology could be implemented as a package of R-project and this work could be applied to other members of the localization and scale family.

[Received: April 2015 — Accepted: February 2016]

References

Allison, P. D. (1995), *Survival Analysis Using the SAS System: A Practical Guide*, Springer-Verlag, New York.

Chang, C. H. & Weissfeld, L. A. (1999), 'Normal aproximation diagnostics for the Cox model', *Biometrics* **55**, 1114–1119.

Cheng, R. & Iles, T. (1983), 'Confidence bands for cumulative distribution functions of continuous random variables', *Technometrics* **25**(1), 77–86.

Cheng, R. & Iles, T. (1988), 'One-sided confidence bands for cumulative distribution functions', *Technometrics* **30**(1), 155–159.

Dorey, F. J., Little, R. & Schenker, N. (1993), 'Multiple imputation for threshold-crossing data with interval censoring', *Statistics in Medicine* **12**, 1589–1603.

Efron, B. (1967), The two sample problem with censored data, Technical report, University of California Press.

Escobar, L. A., Hong, Y. & Meeker, W. Q. (2009), 'Simultaneous confidence bands and regions for log-location-scale distributions with censored data', *Journal of Statistical Planning and Inference* **139**(9), 3231–3245.

Gentleman, R. & Vandal, A. C. (2001), 'Computational algorithms for censored-data problems using intersection graphs', *Journal of Computational and Graphical Statistics* **10**, 403–421.

Jeng, S. & Meeker, W. Q. (2001), 'Parametric simultaneous confidence bands for cumulative distributions from censored data', *Technometrics* **43**(4), 450–461.

Joly, P. & Commenges, D. (1999), 'A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to AIDS', *Biometrics* **55**, 887–890.

Kaplan, E. L. & Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American statistical association* **53**, 457–481.

Lawless, J. & Babineau, D. (2006), 'Models for interval censoring and simulation-based inference for lifetime distributions', *Biometrika* **93**, 671–686.

Meeker, W. & Escobar, L. (1992), 'Assessing influence in regression analysis with censored data', *Biometrics* **48**, 507–528.

Nair, V. N. (1984), 'Confidence bands for survival functions with censored data: A comparative study', *Technometrics* **46**(3), 265–275.

Nelder, J. & Mead, R. (1965), 'A simplex method for function minimization', *Computer Journal* **7**, 308–313.

Odell, P., Anderson, K. & D'Agostinho, R. (1992), 'Maximum likelihood estimation for interval censored data using a Weibull based accelerated failure time model', *Biometrics* **48**, 951–959.

Peto, R. (1973), 'Experimental survival curves for interval-censored data', *Journal of the Royal Statistical Society, Series C* **22**, 86–91.

Rojas, A., Diaz, F. J., Calvo, E., Salazar, J. C., Iglesias, A., Mantilla, R. D. & Anaya, J. M. (2009), 'Familial disease, the HLA-DRB1 shared epitope and anti-CCP antibodies influence time at appearance of substantial joint damage in rheumatoid arthritis', *Journal of Autoimmunity* **32**, 64–69.

Rosales, L. F. & Salazar, J. C. (2006), Estimaciones de funciones de intensidad en un modelo de 3 estados en presencia de doble censura, Master's in statistics, Universidad Nacional De Colombia, Sede Medellín.

Rucker, G. & Messerer, D. (1988), 'Remission duration: an example of interval-censored observation', *Statistics in Medicine* **7**, 1139–1145.

Turnbull, B. W. (1974), 'Nonparametric estimation of a survivorship function with doubly censored data', *Journal of the American statistical association* **69**, 169–173.

Turnbull, B. W. (1976), 'The empirical distribution function with arbitrarily grouped censored and truncated data', *Journal of the Royal Statistical Society, Series B* **38**, 290–295.

Waller, L. A. & Turnbull, B. W. (1992), 'Probability Plotting with censored data', *The American Statistician* **46**, 5–12.