A flexible bimodal model with long-term survivors and different regression structures

GENTIL RAMIRES, Thiago; Ortega, Edwin M. M.; Lemonte, Artur J.; HENS, Niel & Cordeiro, Gauss M. (2020) A flexible bimodal model with long-term survivors and different regression structures. In: COMMUNICATIONS IN STATISTICS-SIMULATION AND COMPUTATION, 49 (10) , p. 2639 -2660.

DOI: 10.1080/03610918.2018.1524902
Handle: http://hdl.handle.net/1942/32902

A flexible bimodal model with long-term survivors and different regression structures

Thiago G. Ramires

Department of Mathematics, Federal University of Tecnology - Paraná, Apucarana, Brazil

Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-Biostat),

University of Hasselt, Belgium;

Edwin M.M. Ortega

Department of Exact Sciences, University of São Paulo, Brazil

Artur J. Lemonte

Department of Statistics, Federal University of Rio Grande do Norte, Brazil

Niel Hens

Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-Biostat),

University of Hasselt, Belgium;

Centre for Health Economic Research and Modelling Infectious Diseases,

Vaccine and Infectious Disease Institute, University of Antwerp, Belgium

Gauss M. Cordeiro

Department of Statistics, Federal University of Pernambuco, Brazil

Abstract

The cure fraction models are useful to model lifetime data with long-term survivors. In this paper, we introduce a flexible cure rate survival model where the model parameters are related to covariates in different regression structures. The regression model allows to model jointly the location, scale and shape effects. The maximum likelihood method is employed to estimate the model parameters. We provide Monte Carlo simulation experiments to verify the performance of the maximum likelihood estimates for different sample sizes and cure rate percentages. Furthermore, some diagnostic measures to assess departures from model assumptions as well as to detect outlying observations are also considered. Finally, applications to real data are presented to show the usefulness of the new cure rate model.

Keywords: Bi-modality; Cure rate models; Parametric inference; Residual analysis; Sensitivity analysis.

1 Introduction

Ramires et al. (2016) introduced the exponentiated log-sinh Cauchy (ELSC) distribution, being its probability density function (PDF) given by

$$f(t) = f(t;\mu,\sigma,\nu,\tau) = \frac{\tau\nu\cosh(w)}{t\sigma\pi[\nu^2\sinh^2(w)+1]}\left\{\frac{1}{2}+\frac{1}{\pi}\arctan\left[\nu\sinh(w)\right]\right\}^{\tau-1}, \quad t>0, \qquad (1)$$

where $w = [\log(t)-\mu]/\sigma$, $\mu \in \mathbb{R}$ and $\sigma > 0$ are the location and scale parameters, respectively, $\nu > 0$ is the symmetry parameter, which characterizes the bi-modality of the distribution, and $\tau > 0$ is the shape (skewness) parameter. The survival function corresponding to (1) is

$$S(t) = S(t;\mu,\sigma,\nu,\tau) = 1-\left\{\frac{1}{2}+\frac{1}{\pi}\arctan\left[\nu\sinh(w)\right]\right\}^{\tau}, \quad t>0. \qquad (2)$$

It is worth emphasizing that the ELSC distribution corresponds to a very flexible lifetime distribution and can take different shapes, including bi-modality (Ramires et al., 2016, § 2). So, it can be used as an interesting alternative to mixture distributions in modeling bimodal data. Additionally, the ELSC hazard function can be constant, decreasing, increasing, upside-down bathtub (unimodal), bathtub and bimodal shaped; see Ramires et al. (2016, § 2) for more details. Theferore, due to the great flexibility of the ELSC model, it thus provides a good alternative to many existing life distributions in modelling positive real data sets.

Models for survival data with a surviving fraction (also known as cure rate models or long-term survival models) occupy an outstanding place in reliability, survival analysis and other areas. Models for survival analysis typically consider that every subject in the study is susceptible to the event under study and will eventually experience such event if follow-up is sufficiently long. However, there are situations when a fraction of individuals are not expected to experience the event of interest, that is, those individuals are cured or not susceptible. Cure rate models for survival data have been used to model time-to-event data for various types of cancers, including breast cancer, non-Hodgkin lymphoma, leukemia, prostate cancer and melanoma. These models have become very popular due to significant progress in treatment therapies leading to enhanced cure rates.

Perhaps the most popular type of cure rate models are the mixture models pioneered by Boag (1949), Berkson and Gage (1952) and further studied by Farewell (1982). The mixture models allow both the cure fraction and the survival function of uncured patients (latency distribution) to depend on covariates. Let M be the indicator denoting that the individual is susceptible $(M = 1)$ or non-susceptible $(M = 0)$; that is, the population is classified in two sub-populations so that an individual either is cured (with probability $0 < p < 1$) or has a proper survival function, $S(t)$ say, with probability $1-p$. The mixture cure rate model can be expressed as

$$S_{\text{pop}}(t) = p + (1-p)S(t|M=1), \quad t>0, \qquad (3)$$

where $S_{\text{pop}}(t)$ is the unconditional survival function for the entire population, $S(t|M=1)$ is the survival function for susceptible individuals, and $p = \Pr(M=0)$ is the probability of cure of an individual. The PDF corresponding to (3) is given by

$$f_{\text{pop}}(t) = -\frac{d\, S_{\text{pop}}(t)}{dt} = (1-p) f(t|M=1), \quad t > 0, \tag{4}$$

where $f(t|M=1)$ is the baseline PDF for the susceptible individuals. Equations (3) and (4) are improper functions, since $S_{\text{pop}}(t)$ is not a proper survival function (i.e. $\lim_{t\to\infty} S_{\text{pop}}(t) > 0$). To introduce the covariate effect, it is quite common to relate the cure rate p to a set of covariates using a logistic link function, for example.

The literature on distributions that accommodate different latent competing causes is rich and growing rapidly. Rodrigues et al. (2009) developed the COM-Poisson cure rate model considering that the number of competing causes of the event of interest follows the Conway-Maxwell Poisson distribution. Ortega et al. (2009) defined the generalized log-gamma regression models with cure fraction to explain/predict the cancer recurrence times. Cancho et al. (2013) proposed a destructive negative binomial cure rate model, where the initial number of competing causes of the event of interest follows a compound negative binomial distribution, and Hashimoto et al. (2014) introduced the Poisson Birbaum-Saunders model with long-term survivors assuming that the number of competing causes of the event of interest follows the Poisson distribution and the time to event has the Birnbaum-Saunders distribution. Rodrigues et al. (2015) studied the relaxed Poisson cure rate model showing an application to cutaneous melanoma data, Ortega et al. (2015) proposed a new cure rate survival regression model for predicting breast carcinoma survival in women who underwent mastectomy, Balakrishnan and Pal (2015) derived an EM algorithm for estimating the parameters of a cure rate model with generalized gamma lifetime, and Balakrishnan et al. (2016) proposed piecewise linear approximations for cure rate models and associated inferential issues, among others.

Although the models studied in the above papers are attractive, they have some limitations. Most of the proposed models are not able to capture the presence of bi-modality in the data; that is, they only support unimodal data. Additionally, another disadvantage is that some of these models only has a regression structure in the cure fraction. In this paper, we consider a novel class of cure rate models based on the ELSC distribution. Assuming a fully parametric setup and considering that the lifetime follows the ELSC distribution, we propose the exponentiated log-sinh Cauchy cure rate (ELSCcr) model, conceived inside a latent competing cause scenario with cure fraction, where there is no information about which cause was responsible, for example, for the individual death or tumor recurrence, but only the minimum lifetime value among all risks is observed and a part of the population is not susceptible to the event of interest. In this new class of cure rate models all model parameters are related to covariates using suitable link functions; that is, the cure rate model allows to model jointly the location, scale and shape effects. Additionally, for the assessment of model adequacy, we develop diagnostic studies to detect possible influential or extreme observations that can cause distortions on the results of the analysis by means of a global influence measure. The residual analysis for the

proposed model is also addressed.

The sections are organized as follows. In Section 2, we introduce the ELSCcr model. Section 3 deals with the ELSCcr regression model. Some strategies to select the best model, residual analysis, goodness-of-fit and global influence measure are addressed in Section 4. Section 5 provides Monte Carlo simulations on the finite sample behavior of the maximum likelihood estimates. Applications to real data are presented in Section 6 to illustrate the new regression model in practice. Finally, we offer some conclusions in Section 7.

2 The ELSC model with long-term survivors

For censored survival times, the presence of an immune proportion of individuals who are not subject to death, failure or relapse may be indicated by a relatively high number of individuals with large censored survival times. Now, we define the ELSCcr model for the possible presence of long-term survivors in the data. To formulate the model, we consider that the population under study is a mixture of susceptible (uncured) individuals, who may experience the event of interest, and non-susceptible (cured) individuals, who will not experience it (Maller and Zhou, 1996).

2.1 Definition

The survival function of the ELSCcr model is defined by assuming that the survival function for susceptible individuals in (3) is given by (2), which leads to

$$S_{\text{pop}}(t) = S_{\text{pop}}(t;\mu,\sigma,\nu,\tau,p) = 1 + (p-1)\left\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu\sinh(w)]\right\}^{\tau}, \tag{5}$$

where $w = [\log(t) - \mu]/\sigma$, $\mu \in \mathbb{R}$, $\sigma > 0$, $\nu > 0$, $\tau > 0$ and $p \in (0,1)$. The PDF corresponding to (5) takes the form

$$f_{\text{pop}}(t) = f_{\text{pop}}(t;\mu,\sigma,\nu,\tau,p) = \frac{(1-p)\tau\nu\cosh(w)}{t\sigma\pi[\nu^2\sinh^2(w)+1]}\left\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu\sinh(w)]\right\}^{\tau-1}. \tag{6}$$

Evidently, the survival function (5) and the PDF (6) do not involve any complicated function. Also, there is no functional relationship between the parameters and they vary freely in the parameter space. Note that even if μ, σ and ν parameters do not have direct interpretation, the model produces estimates of survival, risk functions more accurate for complex behaviors. The hazard function of the ELSCcr model is given by $h_{\text{pop}}(t) = f_{\text{pop}}(t)/S_{\text{pop}}(t)$. If a random variable T has PDF (6), then the notation used is $T \sim \text{ELSCcr}(\mu,\sigma,\nu,\tau,p)$.

Plots of the ELSCcr survival and hazard functions for selected parameter values are displayed in Figures 1 and 2, respectively. These figures reveal clearly the great flexibility of the ELSCcr model. Note that the ELSCcr hazard function can take different shapes depending on the parameter values, which has decreasing, unimodal, bimodal and bathtub-shaped forms.

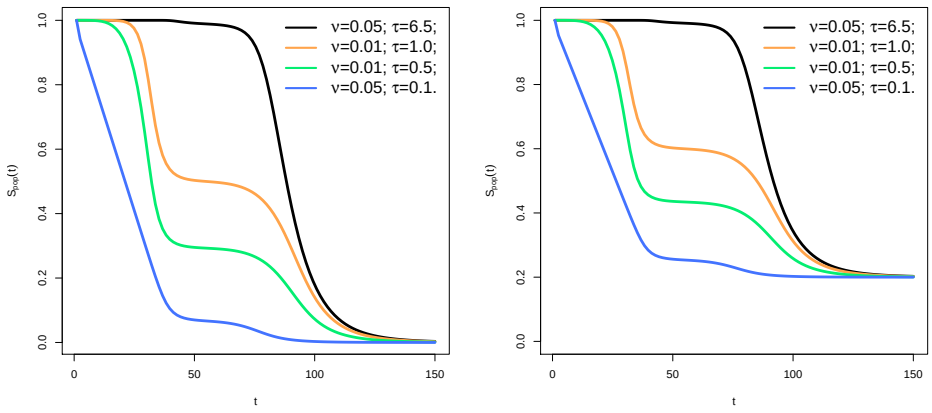

Figure 1: The ELSCcr survival function when $\mu = 4$, $\sigma = 0.1$ and different values of ν and τ for $p = 0$ (left) and $p = 0.2$ (right).

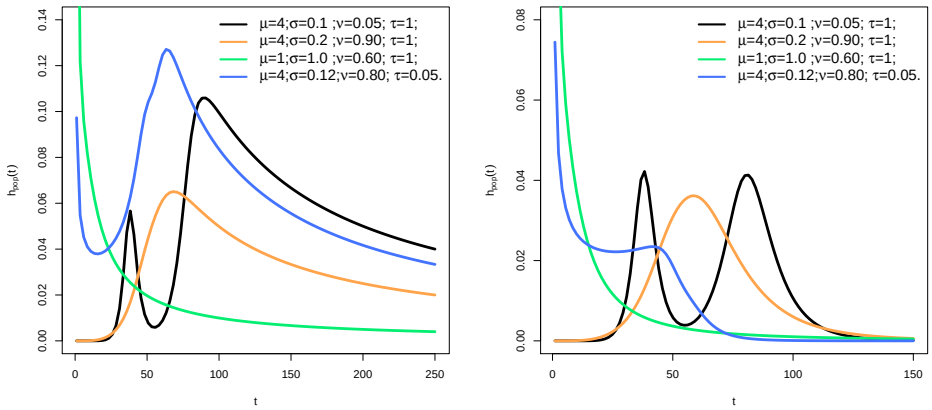

Figure 2: The ELSCcr hazard function for different values of μ, σ, ν and τ, and $p = 0$ (left) and $p = 0.2$ (right).

2.2 Related models

Let $T \sim \text{ELSCcr}(\mu, \sigma, \nu, \tau, p)$ be a random variable having PDF (6). Some sub-models and related distributions are listed in Tables 1 and 2. Note that when $p \neq 0$, the special models presented in both tables below are extended to distributions with cure rate, e.g., when $p \neq 0$, $\sigma = 1$ and $\tau = 1$ (see Table 1, last line), we obtain the folded Cauchy cure rate (FCcr) model; and so on.

Table 1: Related distributions.

Distribution	μ	σ	ν	τ	p	Reference
Exponentiated log-sinh Cauchy (ELSC)	μ	σ	ν	τ	0	Ramires et al. (2016)
Log-sinh Cauchy (LSC)	μ	σ	ν	1	0	Ramires et al. (2016)
Folded Cauchy (FC)	μ	1	ν	1	0	Johnson et al. (1994)

Table 2: Related distributions for $Y = \log(T)$.

Distribution	μ	σ	ν	τ	p	Reference
Exponentiated sinh Cauchy (ESC)	μ	σ	ν	τ	0	Cooray (2013)
Sinh Cauchy (SC)	μ	σ	ν	1	0	Cooray (2013)
Hyperbolic secant (HS)	μ	σ	1	1	0	Talacko (1956)

3 The ELSCcr regression model

In many applications of long term survival models, the cure rate plays an essential role that can be explained by explanatory variables. For example, in medical problems, the lifetimes and the cure rate are affected by the cholesterol level, blood pressure, weight, among others. Parametric models to estimate univariate survival functions for censored data are widely used. Recently, several regression models for long-term survivors have been proposed in the literature, as mentioned in Section 1. In general, these models assume only that the cure rate (p) and/or location (μ) parameters are related to explanatory variables. A disadvantage of the class of location models is that the variance, skewness, bi-modality, kurtosis and other parameters are not modelled explicitly in terms of explanatory variables. As an alternative, we can relate to covariates not only the location parameter but all parameters of the conditional distribution of T. So, we can provide systematic components (i.e. parametric functions of explanatory variables) for all parameters, similarly to the GAMLSS framework (Rigby and Stasinopoulos, 2005).

3.1 Parametric model

Let $T_1, \ldots, T_n$ be n independent positive random variables with T_i having PDF (6) with $\mu = \mu_i$, $\sigma = \sigma_i$, $\nu = \nu_i$, $\tau = \tau_i$ and $p = p_i$, for $i = 1, \ldots, n$; that is, $T_i \sim \text{ELSCcr}(\mu_i, \sigma_i, \nu_i, \tau_i, p_i)$. Let $\boldsymbol{\mu} =$

$(\mu_1, \ldots, \mu_n)^\top$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)^\top$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^\top$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^\top$ and $\boldsymbol{p} = (p_1, \ldots, p_n)^\top$. Suppose that these parameters satisfy the following functional relations:

$$g_1(\mu_i) = \boldsymbol{x}_{i,1}^\top \boldsymbol{\beta}_1, \quad g_2(\sigma_i) = \boldsymbol{x}_{i,2}^\top \boldsymbol{\beta}_2, \quad g_3(\nu_i) = \boldsymbol{x}_{i,3}^\top \boldsymbol{\beta}_3, \quad g_4(\tau_i) = \boldsymbol{x}_{i,4}^\top \boldsymbol{\beta}_4, \quad g_5(p_i) = \boldsymbol{x}_{i,5}^\top \boldsymbol{\beta}_5, \quad (7)$$

where $\boldsymbol{x}_{i,1}^\top = (1, x_{i1,1}, \ldots, x_{im_1,1})$, $\boldsymbol{x}_{i,2}^\top = (1, x_{i1,2}, \ldots, x_{im_2,2})$, $\boldsymbol{x}_{i,3}^\top = (1, x_{i1,3}, \ldots, x_{im_3,3})$, $\boldsymbol{x}_{i,4}^\top = (1, x_{i1,4}, \ldots, x_{im_4,4})$ and $\boldsymbol{x}_{i,5}^\top = (1, x_{i1,5}, \ldots, x_{im_5,5})$ are vectors of known explanatory variables of dimensions $m_1 + 1$, $m_2 + 1$, $m_3 + 1$, $m_4 + 1$ and $m_5 + 1$, respectively, $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{m_k k})^\top$ is a parameter vector of length $m_k + 1$, and $g_k(\cdot)$ denotes injective and twice continuously differentiable monotonic link functions, for $k = 1, 2, 3, 4$ and 5. So, the total number of parameters to be estimated is given by $m = m_1 + m_2 + m_3 + m_4 + m_5 + 5$, where is assumed that $m < n$. It is also assumed that $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_4$ and $\boldsymbol{\beta}_5$ are functionally independent.

From now on, we shall consider the identity link function for $g_1(\cdot)$, and the logarithmic link function for $g_2(\cdot)$, $g_3(\cdot)$ and $g_4(\cdot)$; that is,

$$\mu_i = \boldsymbol{x}_{i,1}^\top \boldsymbol{\beta}_1, \qquad \log(\sigma_i) = \boldsymbol{x}_{i,2}^\top \boldsymbol{\beta}_2, \qquad \log(\nu_i) = \boldsymbol{x}_{i,3}^\top \boldsymbol{\beta}_3, \qquad \log(\tau_i) = \boldsymbol{x}_{i,4}^\top \boldsymbol{\beta}_4.$$

It is worth emphasizing that the estimations of the proportion of cure is very important. The vast majority of researchers adopt the logit link function for relating the cure fraction to covariates, since the logit link allows for a simple representation of the odds ratio, which aids interpretation of the result. However, this popular link does not always provide the best fit for a given data set. In this paper, in addition to the logit link, which is quite usual in long-term survivors, we shall consider the complementary log-log, log-log and probit links for relating the cure fraction to covariates. These links are:

- Logit: $p_i = \dfrac{\exp(\boldsymbol{x}_{i,5}^\top \boldsymbol{\beta}_5)}{1 + \exp(\boldsymbol{x}_{i,5}^\top \boldsymbol{\beta}_5)}$;
- Complementary log-log: $p_i = 1 - \exp[-\exp(\boldsymbol{x}_{i,5}^\top \boldsymbol{\beta}_5)]$;
- Log-log: $p_i = \exp[-\exp(-\boldsymbol{x}_{i,5}^\top \boldsymbol{\beta}_5)]$;
- Probit: $p_i = \Phi(\boldsymbol{x}_{i,5}^\top \boldsymbol{\beta}_5)$;

and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

3.2 Likelihood-based inference

Consider a sample of n independent observations $t_1, \ldots, t_n$, where the response variable y_i corresponds to the observed lifetime or censoring time for the ith individual. We consider non-informative censoring and that the observed lifetimes and censoring times are independent. Let D and C be the

sets of individuals for which y_i is the lifetime or censoring, respectively. The log-likelihood function under non-informative censoring for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top, \boldsymbol{\beta}_4^\top, \boldsymbol{\beta}_5^\top)^\top$ is given by

$$\begin{aligned}\ell(\boldsymbol{\theta}) = \sum_{i\in D}\Big\{&\log(1-p_i) + \log(\tau_i) + \log(\nu_i) - \log(\sigma_i) \\ &- \log(t_i\pi) \log[\cosh(w_i)] - \log\left[1 + \nu_i^2\, \sinh^2(w_i)\right]\Big\} \\ &+ \sum_{i\in D}(\tau_i - 1)\log\left\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu_i \sinh(w_i)]\right\} \\ &+ \sum_{i\in C}\log\left(1 + (p_i - 1)\left\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu_i \sinh(w_i)]\right\}^{\tau_i}\right),\end{aligned} \tag{8}$$

where $w_i = [\log(t_i) - \mu_i]/\sigma_i$. The maximum likelihood (ML) estimates of the unknown parameters are obtained by maximizing the log-likelihood function in (8) with respect to $\boldsymbol{\theta}$. We make some assumptions (Cox and Hinkley, 1974, Ch. 9) on the behavior of $\ell(\boldsymbol{\theta})$ as the sample size n approaches infinity, such as the regularity of the first two derivatives of $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and the existence and uniqueness of the ML estimate of $\boldsymbol{\theta}$.

The score functions for the parameters are given by

$$\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_1} = \boldsymbol{X}_1^\top\boldsymbol{\Omega}_1\boldsymbol{s}_1, \quad \frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_2} = \boldsymbol{X}_1^\top\boldsymbol{\Omega}_2\boldsymbol{s}_2, \quad \frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_3} = \boldsymbol{X}_3^\top\boldsymbol{\Omega}_3\boldsymbol{s}_3,$$

$$\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_4} = \boldsymbol{X}_4^\top\boldsymbol{\Omega}_4\boldsymbol{s}_4, \quad \frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_5} = \boldsymbol{X}_5^\top\boldsymbol{\Omega}_5\boldsymbol{s}_5,$$

where $\boldsymbol{X}_k = (\boldsymbol{x}_{1,k}, \ldots, \boldsymbol{x}_{n,k})^\top$,, for $k = 1, 2, 3, 4, 5$, are known model matrices of full column rank, $\boldsymbol{\Omega}_1 = \boldsymbol{I}_n$ (the identity matrix of order n), $\boldsymbol{\Omega}_2 = \text{diag}\{\sigma_1, \ldots, \sigma_n\}$, $\boldsymbol{\Omega}_3 = \text{diag}\{\nu_1, \ldots, \nu_n\}$, $\boldsymbol{\Omega}_4 = \text{diag}\{\tau_1, \ldots, \tau_n\}$, $\boldsymbol{\Omega}_5 = \text{diag}\{p_1(1-p_1), \ldots, p_n(1-p_n)\}$ for the logit link function, $\boldsymbol{\Omega}_5 = \text{diag}\{(1-p_1)\exp(\boldsymbol{x}_{1,5}^\top\boldsymbol{\beta}_5), \ldots, (1-p_n)\exp(\boldsymbol{x}_{n,5}^\top\boldsymbol{\beta}_5)\}$ for the complementary log-log link function, $\boldsymbol{\Omega}_5 = \text{diag}\{p_1\exp(\boldsymbol{x}_{1,5}^\top\boldsymbol{\beta}_5), \ldots, p_n\exp(\boldsymbol{x}_{n,5}^\top\boldsymbol{\beta}_5)\}$ for the log-log link function, and $\boldsymbol{\Omega}_5 = \text{diag}\{\phi(\boldsymbol{x}_{1,5}^\top\boldsymbol{\beta}_5), \ldots, \phi(\boldsymbol{x}_{n,5}^\top\boldsymbol{\beta}_5)\}$ for the probit link function, where $\phi(\cdot)$ denotes the standard normal PDF. Also, $\boldsymbol{s}_1 = (s_{11}, \ldots, s_{n1})^\top$, $\boldsymbol{s}_2 = (s_{12}, \ldots, s_{n2})^\top$, $\boldsymbol{s}_3 = (s_{13}, \ldots, s_{n3})^\top$, $\boldsymbol{s}_4 = (s_{14}, \ldots, s_{n4})^\top$ and $\boldsymbol{s}_5 = (s_{15}, \ldots, s_{n5})^\top$, and

$$s_{i1} = \begin{cases} \dfrac{\nu_i^2\sinh(2w_i)}{\sigma_i\, K_i} - \dfrac{\tanh(w_i)}{\sigma_i} - (\tau_i - 1)\dfrac{\nu_i\cosh(w_i)}{\pi\sigma_i J_i K_i}, & i \in D, \\ -\dfrac{(p_i-1)\tau_i\nu_i\cosh(w_i)J_i^{\tau_i-1}}{\pi\sigma_i K_i[1 + (p_i-1)\, J_i^{\tau_i}]}, & i \in C, \end{cases}$$

$$s_{i2} = \begin{cases} \dfrac{1 - w_i\tanh(w_i)}{\sigma_i} + \dfrac{\nu_i^2 w_i}{\sigma_i K_i}\sinh(2w_i) - (\tau_i - 1)\dfrac{\nu_i w_i\cosh(w_i)}{\pi\sigma_i J_i K_i}, & i \in D, \\ -\dfrac{\tau_i\nu_i w_i J^{\tau_i-1}\cosh(w_i)}{\pi\sigma_i K_i[1 + (p_i-1)J_i^{\tau_i}]}, & i \in C, \end{cases}$$

$$s_{i3} = \begin{cases} \dfrac{1}{\nu_i} - \dfrac{2\nu_i\sinh^2(w_i)}{K_i} + (\tau_i - 1)\dfrac{\sinh(w_i)}{\pi J_i K_i}, & i \in D, \\ \dfrac{(p_i-1)\tau_i J_i^{\tau_i-1}\sinh(w_i)}{\pi K_i[1 + (p_i-1)J_i^{\tau_i}]}, & i \in C, \end{cases}$$

$$s_{i4} = \begin{cases} \dfrac{1}{\tau_i} + \log(J_i), & i \in D, \\ \dfrac{(p_i - 1)J_i^{\tau_i}\log(J_i)}{1 + (p_i - 1)J_i^{\tau_i}}, & i \in C, \end{cases}$$

$$s_{i5} = \begin{cases} -\dfrac{1}{1 - p_i}, & i \in D, \\ \dfrac{J_i^{\tau_i}}{1 + (p_i - 1)J_i^{\tau_i}}, & i \in C, \end{cases}$$

where $J_i = 1/2 + \pi^{-1}\arctan[\nu_i \sinh(w_i)]$ and $K_i = \nu_i^2 \sinh^2(w_i) + 1$.

The ML estimate $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}_1^\top, \widehat{\boldsymbol{\beta}}_2^\top, \widehat{\boldsymbol{\beta}}_3^\top, \widehat{\boldsymbol{\beta}}_4^\top, \widehat{\boldsymbol{\beta}}_5^\top)^\top$ of $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top, \boldsymbol{\beta}_4^\top, \boldsymbol{\beta}_5^\top)^\top$ can be obtained by solving simultaneously the likelihood equations

$$\left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_1}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}, \quad \left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_2}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}, \quad \left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_3}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}, \quad \left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_4}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}, \quad \left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_5}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}.$$

There is no closed-form expression for the ML estimate $\widehat{\boldsymbol{\theta}}$ and its computation has to be performed numerically using a nonlinear optimization algorithm. The Newton-Raphson iterative technique or the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton nonlinear optimization algorithm could be applied to solve the likelihood equations and obtain $\widehat{\boldsymbol{\theta}}$ numerically. For example, the numerical maximization can be performed in the R software (R Development Core Team, 2016) by using the `optim`, and the `manipulate` package can be used to define initial values for the model parameters.

The covariances of the ML estimates can be obtained using the Hessian matrix. Under standard regularity conditions, confidence intervals and hypothesis tests can be conducted based on the large sample distribution of the ML estimator, which is multivariate normal with covariance matrix given by the inverse of the expected information matrix, i.e. $\widehat{\boldsymbol{\theta}} \sim \mathcal{N}_m(\boldsymbol{\theta}, \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1})$, where the asymptotic covariance matrix is given by $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$, $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = -\mathrm{E}(\ddot{\boldsymbol{L}}_{\theta\theta})$ and $\ddot{\boldsymbol{L}}_{\theta\theta} = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$. Since it is not possible to compute the expected information matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ due to the censored observations (censoring is random and noninformative), we can use the matrix of second derivatives $-\ddot{\boldsymbol{L}}_{\theta\theta}/n$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ to estimate $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Additionally, the observed information matrix used for computing asymptotic confidence intervals for the model parameter can be determined numerically from standard maximization routines, which now provide the observed information matrix as part of their output; e.g., one can use the R functions `optim` or `nlm`, the Ox function `MaxBFGS`, the SAS procedure `NLMixed`, among others, to compute the observed information matrix numerically.

3.3 Selecting explanatory variables and link functions

For the ELSCcr regression model, the selection of covariates for all parameters is performed using a stepwise procedure. There are many different strategies that could be applied for selecting covariates which are related to the parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$, $\boldsymbol{\tau}$ and $\boldsymbol{p}$. Here, we adopt a modification of the strategy described by Voudouris et al. (2012). Let χ be the selection of all terms available for consideration, where χ contains the linear terms. Then, for all terms in χ and for fixed link functions, the strategy consists in two steps. In the first step, we adopt a forward selection procedure to select an appropriate model for $\boldsymbol{\mu}$ based on the AIC criterion, with $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$, $\boldsymbol{\tau}$ and $\boldsymbol{p}$ fitted as constants. After that, we repeat the same procedure to select the model for $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$, $\boldsymbol{\tau}$ and $\boldsymbol{p}$, respectively, using the models already obtained in the previous steps as constants. For the second step, we perform a backward selection procedure to choose an appropriate model for $\boldsymbol{\tau}$, with $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{p}$ fitted as

constants and repeat this procedure for $\boldsymbol{\nu}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$, respectively. At the end of the steps described above, the final model may contain different subsets from χ for $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$. On the other hand, the choice of the link function for the cure fraction can be done using the AIC statistic, or can also be fixed to facilitate interpretation of the parameters.

4 Goodness-of-fit, diagnostics and influence measures

There exists a variety of methodologies to compare several competing models for a given data set and select the one that provides the best fit to the data. The selection of the appropriate distribution is performed in two stages: the fitting stage and the diagnostic stage. In the first stage, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used to compare different fitted models. The diagnostic stage involves the use of residual plots to study departures from the error assumption and the presence of outliers. In the diagnostic stage, we can use influence measures to find those models most affected by atypical observations.

4.1 Choosing the best model

The criteria used for the comparison of the models are $AIC = GD + 2 \times df$ and $BIC = GD + \log(n) \times df$, where $GD = -2\ell(\widehat{\boldsymbol{\theta}})$ represents the global deviance, $\ell(\widehat{\boldsymbol{\theta}})$ is the maximized log-likelihood function and df is the total effective degrees of freedom of the fitted model. The model with the smallest value of such criteria is then selected. The AIC and BIC statistics are asymptotically justified for predicting the goodness-of-fit to the current data, that is, approximations to the average predictive error.

4.2 Diagnostic and influence analysis

Since regression models are sensitive to the underlying model assumptions, generally performing a sensitivity analysis is strongly advisable. In order to study departures from the error assumption and the presence of outliers, we can use the normalized randomized quantile residuals (Dunn and Smyth, 1996). The (normalized randomized quantile) residuals are given by

$$r_{q,i} = \Phi^{-1}(u_i), \quad i = 1, \ldots, n.$$

where $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution, $u_i = 1 - S(y_i|\widehat{\boldsymbol{\theta}})$ and

$$S(y_i|\widehat{\boldsymbol{\theta}}) = 1 + (\hat{p}_i - 1)\left\{\frac{1}{2} + \frac{1}{\pi}\arctan[\widehat{\nu}_i \sinh(\hat{w}_i)]\right\}^{\widehat{\tau}_i},$$

where $\widehat{w}_i = [\log(y_i) - \widehat{\mu}_i]/\widehat{\sigma}_i$, $\widehat{\mu}_i = \boldsymbol{x}_{i,1}^\top \widehat{\boldsymbol{\beta}}_1$, $\widehat{\sigma}_i = \exp(\boldsymbol{x}_{i,2}^\top \widehat{\boldsymbol{\beta}}_2)$, $\widehat{\nu}_i = \exp(\boldsymbol{x}_{i,3}^\top \widehat{\boldsymbol{\beta}}_3)$, $\widehat{\tau}_i = \exp(\boldsymbol{x}_{i,4}^\top \widehat{\boldsymbol{\beta}}_4)$ and $\widehat{p}_i = g_5^{-1}(\boldsymbol{x}_{i,5}^\top \widehat{\boldsymbol{\beta}}_5)$, for $i = 1, \ldots, n$. For censored response variables, u_i is a random value from the uniform distribution on the interval $[1 - S(y_i|\widehat{\boldsymbol{\theta}}), 1]$. The true normalized randomized quantile residuals have a standard normal distribution if the model is correct (Dunn and Smyth, 1996).

The most known perturbation schemes are based on case-deletion (Cook and Weisberg, 1982), where the effects of completely removing cases from the analysis are studied. In the following, a quantity with subscript "$(-i)$" refers to the original quantity with the ith case deleted. So, the log-likelihood function

is denoted by $\ell_{(-i)}(\boldsymbol{\theta})$. Let $\widehat{\boldsymbol{\theta}}_{(-i)} = (\widehat{\boldsymbol{\beta}}_{1(-i)}^\top, \widehat{\boldsymbol{\beta}}_{2(-i)}^\top, \widehat{\boldsymbol{\beta}}_{3(-i)}^\top, \widehat{\boldsymbol{\beta}}_{4(-i)}^\top, \widehat{\boldsymbol{\beta}}_{5(-i)}^\top)^\top$ be the ML estimate of $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top, \boldsymbol{\beta}_4^\top, \boldsymbol{\beta}_5^\top)^\top$ obtained from the maximization of $\ell_{(-i)}(\boldsymbol{\theta})$. To assess the influence of the ith case on the ML estimates, the basic idea is to compare the difference between $\widehat{\boldsymbol{\theta}}_{(-i)}$ and $\widehat{\boldsymbol{\theta}}$. If deletion of a case seriously influences the estimates, for example, changing the inference, more attention should be given to that case. Hence, if $\widehat{\boldsymbol{\theta}}_{(-i)}$ is far from $\widehat{\boldsymbol{\theta}}$, then the ith case is regarded as an influential observation. A popular quatity to measure the difference between $\widehat{\boldsymbol{\theta}}_{(-i)}$ and $\widehat{\boldsymbol{\theta}}$ is the log-likelihood distance defined by

$$LD_i(\boldsymbol{\theta}) = 2\left[\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_{(-i)})\right], \quad i = 1, \ldots, n.$$

We would like to point out that for a specific data set and model, the log-likelihood function can potentially have multiple local maxima and hence we suggest the ML estimate $\widehat{\boldsymbol{\theta}}$ as initial trial vector to obtain the estimate $\widehat{\boldsymbol{\theta}}_{(-i)}$.

5 Monte Carlo experiments

The quantile function of $T \sim \text{ELSC}(\mu, \sigma, \nu, \tau)$ takes the form

$$Q(u) = \exp\left(\mu + \sigma \operatorname{arcsinh}\left\{\frac{1}{\nu}\tan\left[\pi\left(u^{1/\tau} - 0.5\right)\right]\right\}\right), \quad u \in (0,1). \tag{9}$$

Equation (9) can be used for simulating ELSC random variables by fixing μ, σ, ν and τ, and making $T = Q(U)$, where U is a uniform random variable in the $(0,1)$ interval. The cured proportion can be generated using the quantile function of another distribution with real support, fixing p and setting the sample size for cured individuals as $n_c = p \times n$. We can also simulate the ELSCcr regression model by setting the parameters using the parametric structure in (7).

Next, we conduct Monte Carlo simulation experiments to assess the finite sample behavior of the ML estimates of the parameters for different sample sizes and cure rate percentages. In the first simulation, we consider the model presented in Section 2 and, in the second simulation, we consider the ELSCcr regression model by modeling all parameters using explanatory variables. In the two simulation studies, the data are generated by taking $n = 50$ and 100, where the lifetimes T are generated from the ELSC distribution using the quantile function in (9), and the censoring times, denoted by C, are randomly generated from the uniform distribution, that is, $C \sim \mathcal{U}(200, 250)$.

The lifetimes considered in each Monte Carlo step are evaluated as $y_i = \min(t_i, c_i)$. For each configuration of n and p, all results are obtained from 5000 Monte Carlo replications. For each replication, we obtain the ML estimates of the parameters and then, after all replications, we determine the average estimates (AEs), biases and means squared errors (MSEs). The simulations are carried out using the R programming language, where the `optim` is used for maximizing the total log-likelihood function in (8).

5.1 Simulation 1: ELSCcr model

We simulate a bimodal ELSCcr distribution, where $\mu = 4, \sigma = 0.2, \nu = 0.1$, $\tau = 1$ and $p = 0, 0.3, 0.5$. The Monte Carlo results are listed in Table 3. They indicate that the MSEs of the ML estimates decay toward zero as the sample size increases, as expected under first-order asymptotic theory. Note that the estimates are quite

stable and, more important, are close to the true values for the sample sizes considered. The codes in R used in this study can be found in `https://git.io/vpvYt`.

Table 3: The AEs, biases and MSEs for the ELSCcr model with $\mu = 4$, $\sigma = 0.2$, $\nu = 0.1$, $\tau = 2$ and $p = 0, 0.3, 0.5$.

		$n = 50$			$n = 100$		
p	Parameter	AE	Bias	MSE	AE	Bias	MSE
0.0	μ	3.994	-0.006	0.003	3.995	-0.005	0.002
	σ	0.182	-0.018	0.001	0.185	-0.015	0.001
	ν	0.080	-0.020	0.002	0.085	-0.015	0.001
	τ	1.044	0.044	0.043	1.040	0.040	0.021
	p	0.000	0.000	0.001	0.000	0.000	0.001
0.3	μ	3.995	-0.005	0.005	3.993	-0.007	0.002
	σ	0.181	-0.019	0.002	0.188	-0.012	0.001
	ν	0.082	-0.018	0.003	0.088	-0.012	0.002
	τ	1.065	0.065	0.073	1.055	0.055	0.034
	p	0.296	-0.004	0.001	0.296	-0.004	0.000
0.5	μ	3.993	-0.007	0.008	3.992	-0.008	0.003
	σ	0.180	-0.020	0.002	0.187	-0.013	0.001
	ν	0.082	-0.018	0.004	0.087	-0.013	0.002
	τ	1.081	0.081	0.108	1.064	0.064	0.047
	p	0.497	-0.003	0.000	0.497	-0.003	0.000

5.2 Simulation 2: ELSCcr regression model for categorical variables

For the ELSCcr regression model, we consider the lifetimes T composed by the lifetimes of two groups, namely T_0 and T_1; that is, we consider a practical scenario in which subjects are divided into two groups, with one group being treated with the drug (treatment group) and the other group being given the placebo (control group). This group category introduces the covariate (x) in our study where we take $x = 1$ for the treatment group and $x = 0$ for the control group. Let us denote the cured proportions for the treatment and control groups by $p1$ and $p0$, respectively. Consider different characteristics for each group, such as location, scale, asymmetry, bi-modality and cured proportion. We take $g_5(\cdot)$ as the logit link function. We consider $T_0 \sim \text{ELSCcr}(3, 0.08, 0.01, 0.60, 0.11)$ and $T_1 \sim \text{ELSCcr}(3.5, 0.22, 0.22, 4.48, 0.26)$. With this configuration, we have

$$\begin{aligned} &\mu_i = 3 + 0.5x_i, \quad \sigma_i = \exp(-2.5 + 1x_i), \quad \nu_i = \exp(-4.5 + 3x_i), \\ &\tau_i = \exp(-0.5 + 2x_i), \quad p_i = \frac{\exp(-2 + 1x_i)}{1 + \exp(-2 + 1x_i)}. \end{aligned} \tag{10}$$

The results are reported in Table 4 and, for visual analysis, we present in Figure 3 the generated and the estimated (considering the AEs given in Table 4) survival functions by considering the two groups represented

by the covariate x and in Figure 4 the boxplots of the estimated standard errors for each parameter and sample size setting. The Monte Carlo simulations reveal that the MSEs of the ML estimates decay toward zero as n increases, as expected under standard asymptotic theory. The AEs tend to be closer to the true parameter values when n increases. This fact supports that the asymptotic normal distribution provides an adequate approximation to the finite sample distribution of the ML estimator. Also, we may note in Figure 4 that the estimated standard errors are smaller according to larger samples. The codes used for this simulation study can be accessed in `https://git.io/vpfeo`.

Table 4: The AEs, biases and MSEs for the ELSCcr regression model with $\beta_{01} = 3$, $\beta_{11} = 0.5$, $\beta_{02} = -2.5$, $\beta_{12} = 1$, $\beta_{03} = -4.5$, $\beta_{13} = 3$, $\beta_{04} = -0.5$, $\beta_{14} = 2$, $\beta_{05} = -2$ and $\beta_{15} = 1$.

	$n = 50$			$n = 100$		
Parameter	AE	Bias	MSE	AE	Bias	MSE
β_{01}	2.992	-0.008	0.001	2.998	-0.002	0.001
β_{11}	0.510	0.010	0.146	0.534	0.034	0.028
β_{02}	-2.588	-0.088	0.039	-2.562	-0.062	0.017
β_{12}	0.977	-0.023	0.086	0.991	-0.009	0.049
β_{03}	-4.990	-0.490	1.250	-4.802	-0.302	0.522
β_{13}	3.520	0.520	1.563	4.410	0.410	0.725
β_{04}	-0.452	0.048	0.089	-0.486	0.014	0.041
β_{14}	2.222	0.222	1.208	2.061	0.061	0.367
β_{05}	-1.842	0.158	0.051	-1.953	0.047	0.014
β_{15}	0.923	-0.077	0.045	0.960	-0.040	0.021

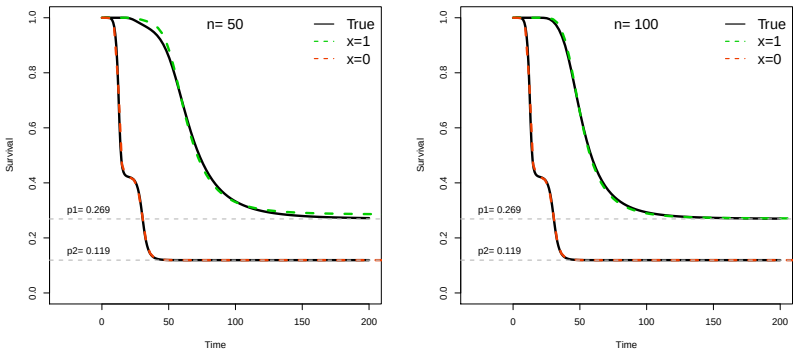

Figure 3: ELSCcr survival functions for $n = 50$ (left) and $n = 100$ (right).

5.3 Simulation 3: ELSCcr regression model for continuous variables

As suggested by an anonymous referee, it is interesting to consider continuous explanatory variables in the simulation experiments. By following this suggestion, we consider another set of simulations considering the

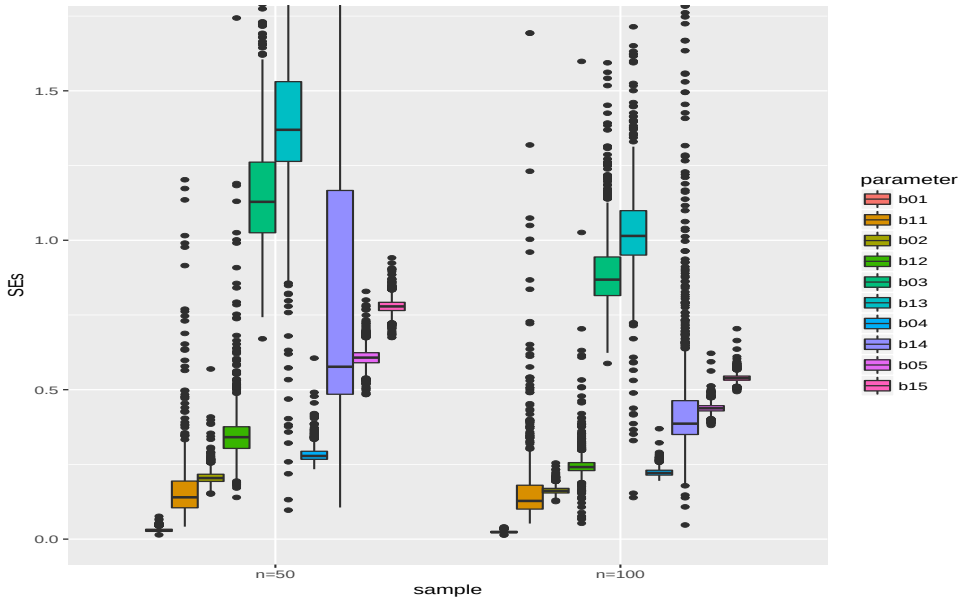

Figure 4: Boxplots of the estimated standard errors.

same regression structure presented in (10), by taking n=200 and 300, where now x is a continuous explanatory variable which assumes the values $x_i = \{0.2, 0.4, 0.6, 0.8, 1\}$. The results are provided in Table 5, which are similar to those presented in Table 4. The estimated standard errors are presented in Figure 5 and, as expected, such estimates are smaller according to larger samples. The results presented in in these last two sections are in accordance with those presented by Ramires et al. (2017) and Alizadeh et al. (2018), which indicate that GAMLSS framework has been increasingly used in the literature. The codes used in this subsection can be found at link `https://git.io/vpf7u`.

Table 5: The AEs, biases and MSEs for the ELSCcr regression model with $\beta_{01} = 3$, $\beta_{11} = 0.5$, $\beta_{02} = -2.5$, $\beta_{12} = 1$, $\beta_{03} = -4.5$, $\beta_{13} = 3$, $\beta_{04} = -0.5$, $\beta_{14} = 2$, $\beta_{05} = -2$ and $\beta_{15} = 1$.

	$n = 200$			$n = 300$		
Parameter	AE	Bias	MSE	AE	Bias	MSE
β_{01}	2.989	-0.011	0.002	2.991	-0.009	0.001
β_{11}	0.532	0.032	0.009	0.529	0.029	0.006
β_{02}	-2.538	-0.038	0.010	-2.537	-0.037	0.007
β_{12}	0.985	-0.015	0.022	0.988	-0.012	0.015
β_{03}	-4.695	-0.195	0.200	-4.696	-0.196	0.156
β_{13}	3.198	0.198	0.337	3.183	0.183	0.267
β_{04}	-0.479	0.021	0.038	-0.488	0.012	0.025
β_{14}	2.034	0.034	0.101	2.032	0.032	0.067
β_{05}	-1.941	0.059	0.021	-1.950	0.050	0.015
β_{15}	0.947	-0.053	0.040	0.947	-0.053	0.030

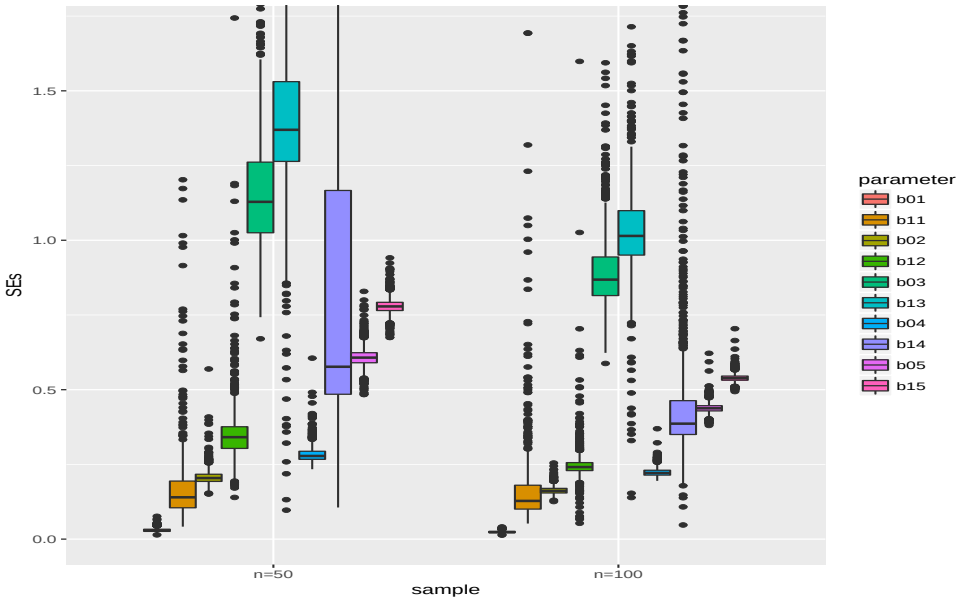

Figure 5: Boxplots of the estimated standard errors.

6 Real data applications

In the following, we provide three applications to real data to demonstrate empirically the flexibility of the ELSCcr model. The first application shows the flexibility of the ELSCcr distribution defined in Section 2 for a sample of independent and identically distributed observations. The second and third applications illustrate empirically the usefulness of the ELSCcr regression model introduced in Section 3. All computations were carried out in the R software by using the `optim` subroutine, and the codes can be downloaded from `https://goo.gl/5Cd8Ug`.

6.1 Calving data

First, we consider the data relative to the ages of the cows at first calving. These data were obtained from the zootechnics records of a Brazilian company engaged in raising beef cattle, located in the states of Bahia and São Paulo. The age at first calving is the main characteristic analyzed, which is an important characteristic for beef cattle breeders due to the fact the faster cows reach reproductive maturity and generating fast return on investment. So, we are interested in modeling the age of the cows (T) at first calving (measured in days).

The sample size in this study is $n = 1326$, where 32.35% of the observations do not present the event of interest (calving) and are thus censored. We assume that $T \sim \text{ELSCcr}(\mu, \sigma, \nu, \tau, p)$. We also compare the results by fitting the LSCcr model, which is a special case of the ELSCcr model when $\tau = 1$, and with the Weibull cure rate model (Weibullcr), which has PDF given by

$$f_{\text{pop}}(t) = (1-p)\frac{\sigma}{\mu}\left(\frac{t}{\mu}\right)^{\sigma-1}\exp(-(t/\mu)^{\sigma}).$$

Table 6 lists the ML estimates of the model parameters and their corresponding standard errors (SEs) in parentheses, and the values of the AIC and BIC statistics for the fitted models. The figures in Table 6 indicate

that the ELSCcr model has the lowest AIC and BIC values, and therefore it could be chosen as the best model. The likelihood ratio (LR) statistic to test the null hypothesis $\mathcal{H}_0 : \tau = 1$ is in accordance with the information criteria (i.e., the observed value of the LR statistic is 126.33 and the associated critical level of the χ^2_1 at 5% is 3.84).

Table 6: ML estimates and the corresponding SEs (in parentheses), and the AIC and BIC statistics; calving data.

Model	μ	σ	ν	τ	p	AIC	BIC
ELSCcr	6.844	0.029	0.023	1.637	0.323	11955.9	11981.9
	(0.002)	(0.001)	(0.003)	(0.069)	(0.013)		
LSCcr	6.855	0.026	0.019		0.320	12080.2	12101.0
	(0.001)	(0.001)	(0.002)		(0.012)		
Weibullcr	1054.340	9.440			0.323	12774.6	12790.1
	(3.067)	(0.192)			(0.018)		

The adequacy of the fitted ELSCcr, LSCcr and Weibullcr models can be noted in Figure 6, which gives the Kaplan-Meier estimate and the estimated survival functions of the fitted distributions for the current data. From this plot, note that the ELSCcr model fits the data adequately and hence can be used to model these data. We also present in Figure 6 the estimated hazard function for the ELSCcr model, where the presence of bi-modality is evident.

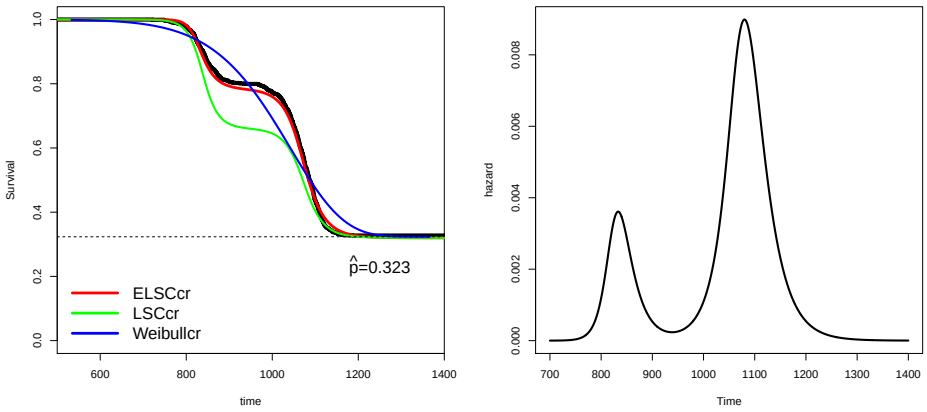

Figure 6: Estimated and empirical survival functions for the ELSCcr, LSCcr and Weibullcr models (left) and the estimated hazard function for the ELSCcr model (right); calving data.

6.2 Gastric cancer data

Gastric cancer is one of the leading causes of cancer-related death and the mucosal resection is accepted as a treatment option for early cases of the disease. It is known that the chemoradiotherapy (CRT) is the standard

treatment used for gastric cancer patients. On the other hand, new technologies to optimize medical decisions and the development of new therapies are of great importance to improve survival in gastric cancer. Jácome et al. (2013) conducted a study in patients with gastric adenocarcinoma who underwent curative resection in which was compared the 3 years overall survival of the two treatments. The study consisted of $n = 201$ patients of different clinical stages, which includes 76 patients that received adjuvant CRT, and 125 that received resection alone. Here, the response variable T refers to the lifetimes (in months) since surgery, and the treatments resection alone and CRT are represented by a dichotomous covariate $x = 0$ and $x = 1$, respectively. We consider censored the lifetimes of the patients who remain alive after the end of the study. These data are provided in Martinez et al. (2013).

Using the steps described in the previous sections to select the additive terms for the different parameters for the ELSCcr regression model, we arrived at the following systematic components:

$$\mu_i = \beta_{01}, \quad \sigma_i = \exp(\beta_{02}), \quad \nu_i = \exp(\beta_{03} + \beta_{13}x_i),$$

$$\tau_i = \exp(\beta_{04} + \beta_{14}x_i), \quad g_5(p_i) = \beta_{05} + \beta_{15}x_i,$$

where $i = 1, \ldots, 201$, $g_5(\cdot)$ can be taken as the logit, complementary log-log, log-log or probit link functions. Table 7 lists the values of the AIC and BIC statistics for the fitted models under different link functions for $g_5(\cdot)$. From this table we have that the log-log link function gives the lowest values of AIC and BIC statistics. Table 8 provides the MLEs, SEs and p-values obtained for the ELSCcr regression model by taking $g_5(\cdot)$ as the log-log link. Note that the parameter β_{15} is not significant at 5%, which indicates that there is no evidence of differences between the population cure fractions considering patients treated by adjuvant CRT and surgery alone.

Table 7: AIC and BIC statistics for the fitted models; gastric cancer data.

Link function	AIC	BIC
Logit	869.4	895.9
Complementary log-log	869.5	896.0
Log-log	869.3	895.7
Probit	869.7	896.2

Table 8: ML estimates, SEs and p-values; gastric cancer data.

Parameter	Estimate	SE	p-value	Parameter	Estimate	SE	p-value
β_{01}	2.994	0.072	< 0.001	β_{04}	-1.995	0.368	< 0.001
β_{02}	-1.575	0.350	< 0.001	β_{14}	1.657	0.266	< 0.001
β_{03}	0.285	0.282	0.157	β_{05}	0.283	0.133	0.017
β_{13}	-1.064	0.522	0.021	β_{15}	0.111	0.258	0.332

To verify the adequacy and the assumptions of the fitted ELSCcr regression model to the data, we present in Figure 7 the index plot for the quantile residuals. Note that the quantile residuals appear satisfactory (random). We also present in Figure 7 the case deletion measure discussed in the previous sections, and no case appears as

a possibly influential observation. Additionally, the Kaplan-Meier estimate and the estimated survival function of the ELSCcr regression model are displayed in Figure 8. In short, we can conclude that the ELSCcr regression model provides a good fit to these data. We also present in Figure 8 the estimated hazard functions, which reveal that the hazard of death is higher in the time immediately after the surgery considering the patients that received the surgery alone. On the other hand, for the patients that received the chemoradiotherapy, the hazard of death has bimodal form with high values at 15 and 27 months after the surgery intervention.

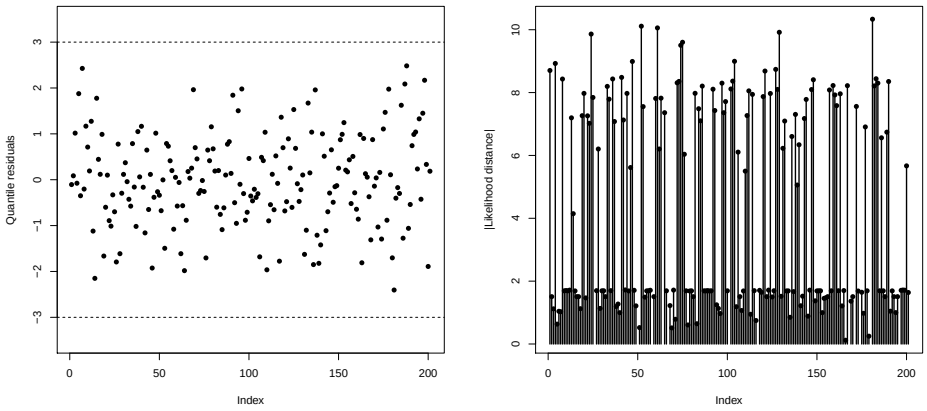

Figure 7: Index plot of the quantile residuals (left), and the absolute values of likelihood distance (right); gastric cancer data.

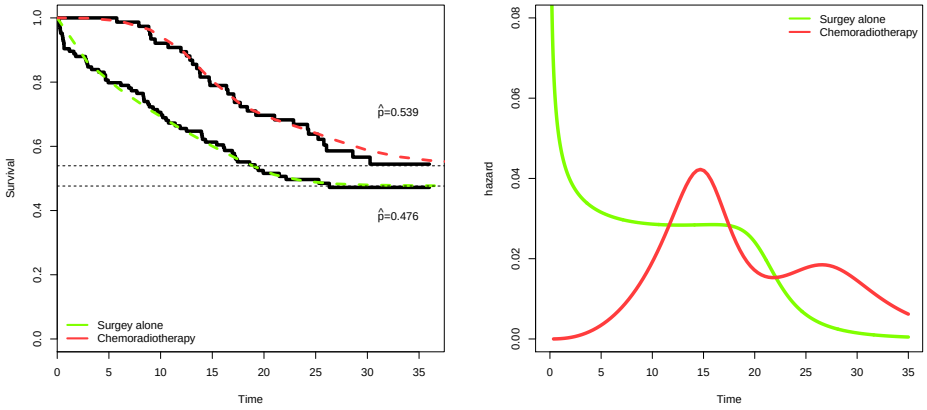

Figure 8: Estimated and empirical survival functions (left), and the estimated hazard functions (right); gastric cancer data.

6.3 Breast cancer data

Recently, several surveys have been developed to identify factors related to breast cancer, which indicates that some conventional clinical factors such as tumor grade, size, surgical margins, among others, are no longer sufficient as prognostic factors. Haque et al. (2012) suggested that breast cancer subtypes are important to consider in treatment decision making. Four main major breast cancer subtypes have been identified, namely Lumial A, Lumial B, Basal and Her2, which are classified using molecular subtyping methods.

To construct the real data set used in this application, we consider five data sets that are available as experimental data packages on `https://www.bioconductor.org/`. Molecular information has been extracted from the phenotype (pData) of the corresponding data set under the Gene Expression Omnibus (GEO). To perform molecular sub-typing, we adopt the SCMOD2 sub-typing algorithms. The steps to construct these data can be found in Gendoo et al. (2015). The final data set consists of $n = 493$ observations containing the lifetime (in months) of patients as well as the breast cancer subtypes, which are represented by dummies variables as follows: Basal ($x_1 = 0$, $x_2 = 0$, $x_3 = 0$), Her2 ($x_1 = 1$, $x_2 = 0$, $x_3 = 0$), Lumial A ($x_1 = 0$, $x_2 = 1$, $x_3 = 0$) and Lumial B ($x_1 = 0$, $x_2 = 0$, $x_3 = 1$).

By using the procedure described in the previous sections for selecting the additive terms for the different parameters, we arrived at the following systematic components:

$$\mu_i = \beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2} + \beta_{31}x_{i3}, \quad \sigma_i = \exp(\beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2} + \beta_{32}x_{i3}),$$

$$\nu_i = \exp(\beta_{03} + \beta_{13}x_{i1} + \beta_{23}x_{i2} + \beta_{33}x_{i3}), \quad \tau_i = \exp(\beta_{04} + \beta_{14}x_{i1} + \beta_{24}x_{i2} + \beta_{34}x_{i3}),$$

$$g_5(p_i) = \beta_{05} + \beta_{15}x_{i1} + \beta_{25}x_{i2} + \beta_{35}x_{i3},$$

where $i = 1, \ldots, 493$. Table 9 lists the values of the AIC and BIC statistics for the fitted models under different link functions for $g_5(\cdot)$. We have that the logit link function gives the lowest values of the AIC and BIC statistics. Table 10 provides the ML estimates, SEs and p-values for the ELSCcr regression model by taking $g_5(\cdot)$ as the logit link function. Note that β_{25} is significative at the 1% level, indicating a difference between the population cure rate fractions of Lumial A and Basal subtypes. We can also note that the subtypes have a significant effect on the location, scale, skewness and bi-modality parameters, so it should be used to obtain accurate estimates.

Table 9: AIC and BIC statistics for the fitted models; breast cancer data.

Link functions	AIC	BIC
Logit	1799.9	1883.9
Complementary log-log	1800.0	1884.0
Log-log	1803.5	1887.5
Probit	1801.1	1885.1

The index plot of the quantile residuals and the likelihood distance are displayed in Figure 9 in order to verify the adequacy and the assumptions of the proposed model. Note that the quantile residuals appear satisfactory (random) and hence the ELSCcr regression model provides a good fit to the data. As can be observed, these plots highlight the observation #447, which corresponds to the lowest value of lifetime for the

Table 10: ML estimates, SEs and p-values; ; breast cancer data.

Parameter	Estimate	SE	p-value	Parameter	Estimate	SE	p-value
β_{01}	4.084	0.078	< 0.001	β_{23}	4.124	0.509	< 0.001
β_{11}	-0.619	0.427	0.074	β_{33}	0.326	0.682	0.316
β_{21}	1.172	0.078	< 0.001	β_{04}	-0.198	0.256	0.221
β_{31}	0.643	0.111	< 0.001	β_{14}	0.660	0.750	0.190
β_{02}	-1.472	0.191	< 0.001	β_{24}	-1.724	0.318	< 0.001
β_{12}	0.366	0.461	0.214	β_{34}	0.019	0.458	0.484
β_{22}	-0.744	0.191	< 0.001	β_{05}	-0.004	0.272	0.494
β_{32}	0.015	0.330	0.482	β_{15}	0.303	0.397	0.223
β_{03}	-2.223	0.509	< 0.001	β_{25}	1.234	0.399	0.001
β_{13}	1.180	0.788	0.067	β_{35}	-0.415	0.627	0.254

Lumial A subtype. Additionally, the adequacy of the fits can also be observed in Figure 10, which presents the empirical and estimated survival function for each breast cancer subtypes. The fitted hazard functions are also given in Figure 11, where we observe bimodal shapes for the Basal, Her2 and Lumial B subtypes. These plots evidence the non-proportionality of the hazard functions, making attractive the use of parametric models for the analysis of these data since they do not consider the assumption of proportional hazards used in the usual semi-parametric Cox model. We can conclude that the ELSCcr regression model yields a good fit for the breast cancer data.

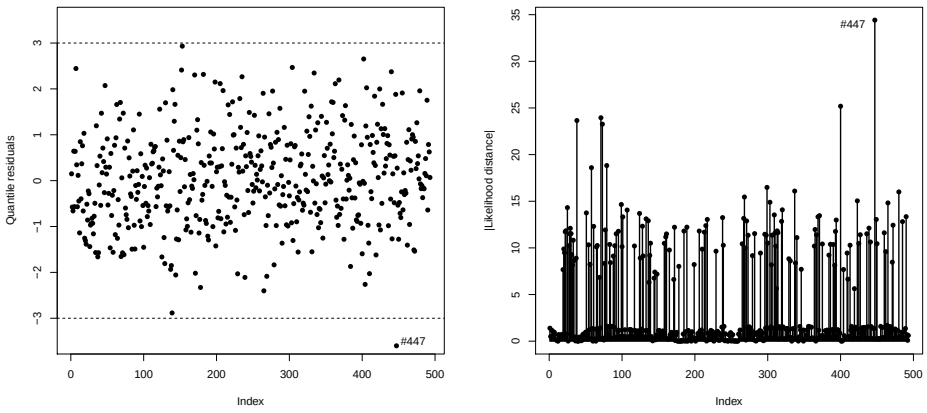

Figure 9: Index plot of the quantile residuals (left), and the absolute values of likelihood distance (right); breast cancer data.

7 Conclusions

We propose the exponentiated log-sinh Cauchy cure rate (ELSCcr) model that can be used as an alternative to mixture distributions in modeling bimodal data with or without the presence of immune proportion of in-

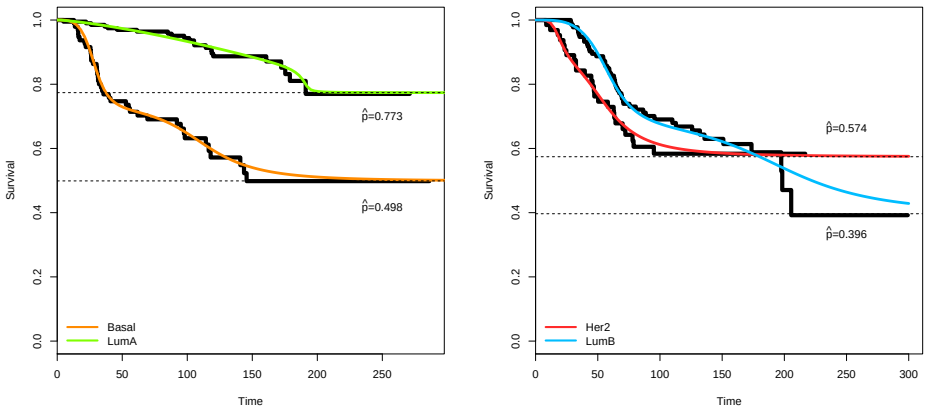

Figure 10: Estimated and empirical survival functions for the Basal and Lumial A subtypes (left), and for Her2 and Lumial B subtypes (right); breast cancer data.

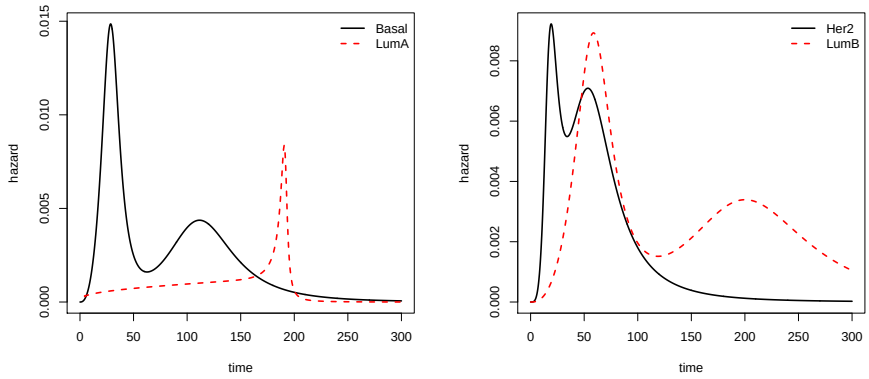

Figure 11: Estimated hazard functions for the Basal and Lumial A subtypes (left), and for the Her2 and Lumial B subtypes (right).

dividuals. We also provide regression structures for all parameters related to location, scale, bi-modality and skewness, which are expressed as linear functions of explanatory variables. We discuss some diagnostic techniques for the ELSCcr regression model, and we also consider residuals for this class of regression models that are very simple to be considered in practical applications. In particular, we highlight the use of normalized quantile residuals to check the adequacy of the new regression model in practical applications. Some numerical experiments reveal that the maximum likelihood estimation procedure works well in estimating the model parameters. Finally, three real data examples prove empirically are presented, showing that the ELSCcr model is very flexible, parsimonious, and a competitive model that deserves to be added to existing distributions in modeling bimodal data.

Acknowledgements

The authors acknowledge the financial support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq/Brazil).

References

Alizadeh, M., Ramires, T.G., MirMostafaee, S.M.T.K., Samizadeh, M., Ortega, E.M.M. (2018). A new useful four-parameter extension of the Gumbel distribution: Properties, regression model and applications using the GAMLSS framework. Communications in Statistics-Simulation and Computation, DOI: 10.1080/03610918.2018.1423691.

Balakrishnan, N., Koutras, M.V., Milienos, F., Pal, S. (2016). Piecewise linear approximations for cure rate models and associated inferential issues. *Methodology and Computing in Applied Probability*, **18**, 937–966.

Balakrishnan, N., Pal, S. (2015). An EM algorithm for the estimation of flexible cure rate model parameters with generalized gamma lifetime and model discrimination using likelihood- and information-based methods. *Computational Statistics*, **30**, 151–189.

Berkson, J., Gage, R.P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**, 501–515.

Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society B*, **11**, 15–53.

Cancho, V.G., Bandyopadhyay, D., Louzada, F., Yiqi, B. (2013). The destructive negative binomial cure rate model with a latent activation scheme. *Statistical Methodology*, **13**, 48–68.

Cook, R.D., Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

Cooray, K. (2013). Exponentiated sinh Cauchy distribution with applications. *Communications in Statistics-Theory and Methods*, **42**, 3838–3852.

Cox, D.R., Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Dunn, P.K., Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.

Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, **38**, 1041–1046.

Gendoo, D.M.A., Ratanasirigulchai, N., Schroder, M., Pare, L., Parker, J.S., Prat, A., Haibe-Kains, B. (2015). genefu: a package for breast cancer gene expression analysis. Retrieved 2016-03-30, from https://bioc.ism.ac.jp/packages/devel/bioc/vignettes/genefu/inst/doc/genefu.PDF

Haque, R., Ahmed, S.A., Inzhakova, G., Shi, J., Avila, C., Polikoff, J., Bernstein, L., Enger, M.S., Press, M.F. (2012). Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. *Cancer Epidemiology Biomarkers & Prevention*, **21**, 1848–1855.

Hashimoto, E.M., Ortega, E.M.M., Cordeiro, G.M., Cancho, V.G. (2014). The Poisson Birnbaum-Saunders model with long-term survivors. *Statistics*, **48**, 1394–1413.

Jácome, A.A.A., Wohnrath, D.R., Neto, C.S., Fregnani, J.H.T.G., Quinto, A.L., Oliveira, A.T.T., Vazquez, V.L., Fava, G., Martinez, E.Z., Santos, J.S. (2013). Effect of adjuvant chemoradiotherapy on overall survival of gastric cancer patients submitted to D2 lymphadenectomy. *Gastric Cancer*, **16**, 233–238.

Johnson, N.L., Kotz, S., Balakrishnan, N. (1994). *Continuous Univariate Distributions*. Wiley, New York.

Maller, R.A., Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, New York.

Martinez, E.Z., Achcar, J.A., Jácome, A.A.A., Santos, J.S. (2013). Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data. *Computer Methods and Programs in Biomedicine*, **112**, 343–355.

Ortega, E.M.M, Cancho, V.G., Paula, G.A. (2009). Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis*, **15**, 79–106.

Ortega, E.M.M, Cordeiro, G.M., Campelo, A.K., Kattan, M.W., Cancho, V.G. (2015). A power series beta Weibull regression model for predicting breast carcinoma. *Statistics in Medicine*, **34**, 1366–1388.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ramires, T.G., Ortega, E.M.M., Cordeiro, G.M., Hens, N. (2016). A bimodal flexible distribution for lifetime data. *Journal of Statistical Computation and Simulation*, **86**, 2450–2470.

Ramires, T.G., Ortega, E.M., Cordeiro, G.M., Paula, G.A., Hens, N. (2017). New regression model with four regression structures and computational aspects. *Communications in Statistics-Simulation and Computation*, DOI: 10.1080/03610918.2017.1332212.

Rigby, R.A., Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society C*, **54**, 507–554.

Rodrigues, J., de Castro, M., Cancho, V.G., Balakrishnan, N. (2009). COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, **139**, 3605–3611.

Rodrigues, J., Cordeiro, G.M., Cancho, V.G., Balakrishnan, N. (2015). Relaxed Poisson cure rate models. *Biometrical Journal*, **58**, 397–415.

Talacko, J. (1956). Perks' distributions and their role in the theory of Wiener's stochastic variables. *Trabajos de Estadística*, **7**, 159–174.

Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J., Stasinopoulos, D. (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, **39**, 1279–1293.