Check for updates

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2018, Vol. 27(11) 3207–3223
© The Author(s) 2017
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280217695344
journals.sagepub.com/home/smm
SAGE

Predicting the cure rate of breast cancer using a new regression model with four regression structures

Thiago G Ramires,[1,2] Gauss M Cordeiro,[3] Michael W Kattan,[4] Niel Hens[2,5] and Edwin MM Ortega[1]

Abstract

Cure fraction models are useful to model lifetime data with long-term survivors. We propose a flexible four-parameter cure rate survival model called the log-sinh Cauchy promotion time model for predicting breast carcinoma survival in women who underwent mastectomy. The model can estimate simultaneously the effects of the explanatory variables on the timing acceleration/deceleration of a given event, the surviving fraction, the heterogeneity, and the possible existence of bimodality in the data. In order to examine the performance of the proposed model, simulations are presented to verify the robust aspects of this flexible class against outlying and influential observations. Furthermore, we determine some diagnostic measures and the one-step approximations of the estimates in the case-deletion model. The new model was implemented in the *generalized additive model for location, scale and shape* package of the R software, which is presented throughout the paper by way of a brief tutorial on its use. The potential of the new regression model to accurately predict breast carcinoma mortality is illustrated using a real data set.

Keywords

Cure rate models, regression models, residual analysis, sensitivity analysis, generalized additive model for location, scale and shape

1 Introduction

Breast cancer, as the name indicates, affects the breasts, which are glands formed by lobes, in turn divided into smaller structures called lobules and ducts. It is the most common malignant tumor among women and the one that causes the most deaths. For example, according to statistics, Brazil had about 576,000 new cases of cancer in 2014–2015, of which over 57,000 were breast cancer. Breast cancer is relatively rare before the age of 35, but above this age, its incidence rises rapidly. However, it is important to remember that not all tumors of the breast are malignant, and that breast cancer can also occur in men, although at a much lower rate. The majority of nodules (or lumps) detected in the breast are benign, but this can only be confirmed through medical tests. Tumors of this size are too small to detect by palpation but are visible in mammograms. Therefore, it is fundamental for all women to be examined by mammography once a year as of the age of 40 years. Breast cancer—and cancer in general—does not have a single cause. Its development is a function of a series of risk factors, some of them modifiable and others not. When diagnosed and treated in the early stage (when the nodule is smaller than 1 cm in diameter), the chances of curing breast cancer are up to 95%. On the other hand, with the advancement of pharmaceutical research, development of new drugs, the chances of a cure as well as the survival times are

[1] Department of Exact Sciences, University of São Paulo, Piracicaba, Brazil
[2] Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-Biostat), University of Hasselt, Hasselt, Belgium
[3] Department of Statistics, Federal University of Pernambuco, Pernambuco, Brazil
[4] Department of Quantitative Health Sciences, Cleveland Clinic, Desk JJN3-01, OH, USA
[5] Centre for Health Economic Research and Modelling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

Corresponding author:
Thiago G Ramires, Department of Exact Sciences, University of São Paulo, 13418-900 Piracicaba, Brazil.
Email: thiagogentil@gmail.com

increasing, requiring a flexible statistical distribution to model such facts. In this study, we address the log-sinh Cauchy promotion time model assuming that part of the population is cured.

Models to accommodate a cured fraction have been widely developed. Models for survival analysis typically assume that all units under study are susceptible to the event and will eventually experience this event if the follow-up is sufficiently long. However, there are situations for which a fraction of individuals is not expected to experience the event of interest; that is, those individuals are cured or insusceptible. Perhaps, the most popular type of cure rate models is the mixture models (MMs) pioneered by Berkson and Gage,[1] Boag[2] and Farewell.[3] MMs allow simultaneously estimating whether the event of interest occurs, which is called incidence, and when it occurs, given that it occurs, which is called latency. The disadvantage of the MMs is that they do not have a biological interpretation. As an alternative to the MMs, Yakovlev and Tsodikov[4] introduced the promotion time cure model, based on a biological context. The main difference between the MMs and promotion time cure models is that in the MMs, the unknown number of causes of the event of interest is assumed to be a binary random variable on $\{0, 1\}$, and in the promotion time cure modeling, this number follows a Poisson distribution. In a biological context, the idea behind these assumptions lies within a latent competing cause structure, in the sense that the event of interest can be the death of a patient or a tumor recurrence, which can happen due to unknown competing causes. If there is no death or tumor recurrence, the patient can be considered cured.

To introduce the promotion time cure models,[4] we consider that $M \sim \text{Poisson}(\tau)$ represents the number of cases for the breast cancer and Z_i denotes the time until the cancer becomes detectable for the ith individual. Given M, the random variables Z_i, for $i = 1, \ldots, M$, are assumed to be independent and identically distributed with a common distribution function $F(z) = 1 - S(z)$ that does not depend on M. The time until the cancer being detected corresponds to the shortest among the M promotion times. Thus, the delay to detectability may be represented by the random variable $T = \{\min Z_i, 0 \leq i \leq M\}$, where $P(Z_0 = 1) = 1$. The resulting survival function for the entire population is

$$S_p(t) = \exp[-\tau F(t)] \tag{1}$$

where $S_p(t)$ is the unconditional survival function of t for the entire population. Note that when $t \to \infty$, $S_p(t) \to \mathrm{e}^{-\tau} = p$, where $0 \leq p \leq 1$ denotes the cured proportion. The probability density function (pdf) corresponding to the survival function ((equation (1)) is given by

$$f_p(t) = \tau f(t) \exp[-\tau F(t)] \tag{2}$$

Note that equation (2) is an improper function, since $S_p(t)$ is not a proper survival function. These latent competing causes M can be assigned to metastasis-component tumor cells left active after an initial treatment.[5] Latent variables represent a theoretical issue and are not observable, so they cannot be measured directly. However, they can be measured by other variables. Genes with low and high expression are significant factors in the lifetime of patients with breast cancer, which may cause lifetimes with bimodal densities.[6] Due to this fact, flexible statistical models are needed to predict as well as correctly identify explanatory variables that may influence the lifetimes of patients diagnosed with breast cancer. In this sense, for modeling a lifetime $T > 0$, the log-sinh Cauchy (LSC) distribution[7] was introduced to accommodate various shapes of skewness, kurtosis and bi-modality. The LSC pdf can be expressed as

$$f(t; \mu, \sigma, \nu) = \frac{\nu}{t\sigma\pi} \frac{\cosh\left(\dfrac{\log(t) - \mu}{\sigma}\right)}{\nu^2 \sinh^2\left(\dfrac{\log(t) - \mu}{\sigma}\right) + 1} \tag{3}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are the location and scale parameters, respectively, and $\nu > 0$ is the symmetry parameter, which characterizes the bi-modality of the distribution. The advantage of the LSC distribution is that it accommodates various shapes of the skewness, kurtosis and bi-modality and can be used as an alternative to mixture distributions in modeling bimodal data. The cumulative distribution function (cdf) corresponding to equation (3) is given by

$$F(t; \mu, \sigma, \nu) = \frac{1}{2} + \frac{1}{\pi} \arctan\left[\nu \sinh\left(\frac{\log(t) - \mu}{\sigma}\right)\right] \tag{4}$$

A standard assumption in regression analysis with censored data is homogeneity of the error variances. Violation of this assumption can have adverse consequences for the efficiency of estimators, so it is important to check for heteroscedasticity whenever it is considered a possibility. In this paper, we propose a general class of regression models with cure fraction, where mean, dispersion, bi-modality and cure fraction parameters vary across observations through regression structures.

The assessment of robustness of the parameter estimates in statistical models has more recently been an important concern. For example, Ortega et al.[8] investigated local influence in generalized log-gamma regression models with cure fraction, Silva et al.[9] adapted global and local influence methods in log-Burr XII regression models with censored data and Hashimoto et al.[10] proposed the log-Burr XII regression model for grouped survival data. The influence diagnostic is an important step in the analysis of a data set as it provides an indication of bad model fitting or of influential observations. The case deletion measures, which consist of studying the impact on the parameter estimates after dropping individual observations, are probably the most employed technique to detect influential observations. We develop a similar methodology to detect influential subjects in the new regression model with long-term survivors.

On the other hand, many researchers have introduced new models in computational packages for ease of use by other researchers. The COM-Poisson cure rate model[11] was introduced in the *generalized additive model for location, scale and shape* (GAMLSS)[12] package of the R software,[13] considering that the number of competing causes of the event of interest follows the Conway–Maxwell–Poisson distribution; some long-term survival models were implemented by taking the Weibull as the parent distribution[5]; the standard mixture Weibull model with a frailty term was also introduced in the GAMLSS package,[14] incorporating heterogeneity of two subpopulations to the event of interest. We set the new model in the GAMLSS package, for which the introduction and all instructions for using are discussed in the following sections.

The paper is organized as follows. In Section 2, we propose the *log-sinh Cauchy promotion time* (LSCp) model by defining the density, cumulative and survival and hazard functions and discuss inferential issues. In Section 3, we introduce the log-sinh Cauchy promotion time regression model, where the parameters can be modeled as function of explanatory variables using the GAMLSS framework. We also discuss inferential issues in this section. Strategies to select the best model, residual analysis, goodness of fit and global influence measure are addressed in Section 4. Section 5 contains methods for generating random values and two Monte Carlo simulations on the finite sample behavior of the maximum likelihood estimates (MLEs). Application to breast cancer data is presented in Section 6 to illustrate the flexibility of the new regression model. Finally, we offer some conclusions in Section 7.

2 The LSCp model

Based on the LSC distribution, we define the LSCp model by inserting equations (3) and (4) in equation (2). The pdf and survival function of the LSCp model are given by

$$f_p(t;\mu,\sigma,\nu,\tau)=\frac{\tau\nu}{t\sigma\pi}\frac{\cosh(w)}{\nu^2\ \sinh^2(w)+1}\exp\Big\{-\frac{\tau}{2}-\frac{\tau}{\pi}\arctan\left[\nu\ \sinh(w)\right]\Big\} \tag{5}$$

and

$$S_p(t;\mu,\sigma,\nu,\tau)=\exp\Big\{-\frac{\tau}{2}-\frac{\tau}{\pi}\arctan\left[\nu\sinh(w)\right]\Big\} \tag{6}$$

respectively, where $w=\frac{\log(t)-\mu}{\sigma}$, $\mu\in\mathbb{R}$ and $\sigma>0$ are the location and scale parameters, respectively, $\nu>0$ is the symmetry parameter, characterizing the bimodality of the distribution, and $\tau>0$ is the cure rate parameter. A random variable having density (equation (5)) is denoted by $T\sim LSCp(\mu,\sigma,\nu,\tau)$. We can omit the dependence on the parameters to simplify notation, for example, $S_p(t)=S_p(t;\mu,\sigma,\nu,\tau)$.

The survival function for non-cured individuals and the hazard rate function (hrf) of the LSCp model are given, respectively, by

$$S(t;\mu,\sigma,\nu,\tau)=\frac{\exp\big\{-\frac{\tau}{2}-\frac{\tau}{\pi}\arctan\left[\nu\ \sinh(w)\right]\big\}-\exp(-\tau)}{1-\exp(-\tau)} \tag{7}$$

and

$$h_p(t;\mu,\sigma,\nu,\tau)=\frac{\tau\nu}{t\sigma\pi}\frac{\cosh(w)}{\nu^2\sinh^2(w)+1} \tag{8}$$

Note that the $h_p(t)$ is multiplicative in τ and $f(t)$; thus, it has the proportional hazard structure. The identifiability between the parameters in cure fraction and those in the time failure distribution for the cure model have been discussed in literature.[15–17] The cure model in equation (1) is identifiable if $F(.)$ is a parametric model.[17]

```
source(‘‘https://goo.gl/gx3t66‘‘)
library(gamlss.cens);library(gamlss)
dLSCp(t,mu,sigma,nu,tau)#pdf
pLSCp(t,mu,sigma,nu,tau)#cdf=1-S(t)
hLSCp(t,mu,sigma,nu,tau)#hrf
```

The functions (5), (6) and (8) are implemented in the R software and can be easily accessed by following the steps in the box displayed above. Plots of the LSCp survival and hazard functions for selected parameter values are displayed in Figures 1 and 2, respectively. Figure 1 reveals clearly the bi-modality and symmetric effects caused by

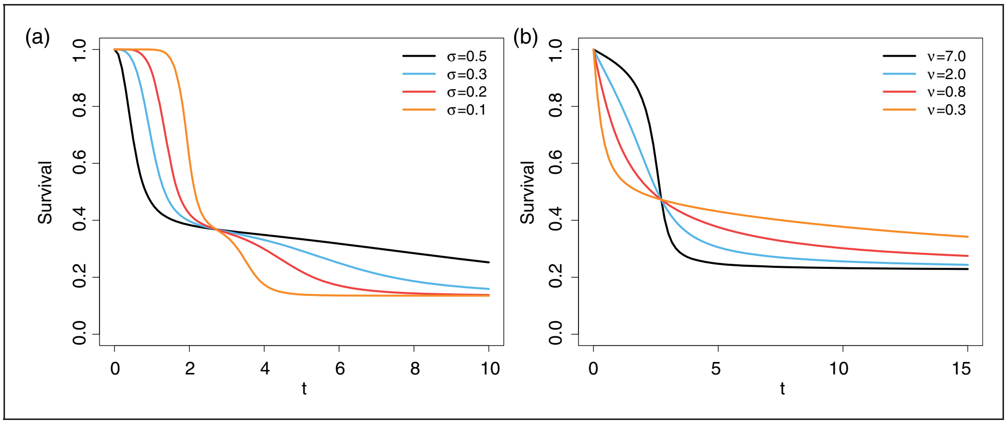

Figure 1. The LSCp survival function when $\mu = 1$ and: (a) For $\nu = 0.1,\ \tau = 2$ and different values of σ; (b) For $\sigma = 1,\ \tau = 1.5$ and different values of ν.

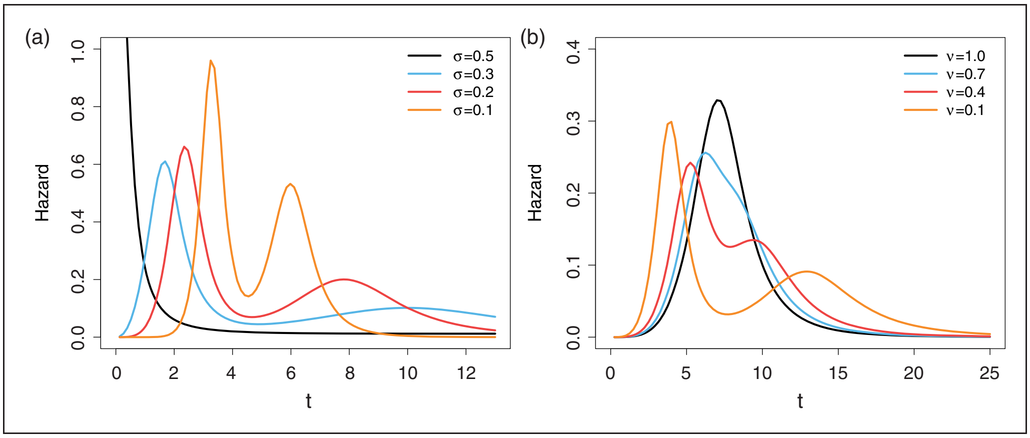

Figure 2. The LSCp hrf for (a) $\mu = 1.5,\ \nu = 0.1,\ \tau = 2$ and different values of σ; (b) $\mu = 2,\ \sigma = 0.2,\ \tau = 1.5$ and different values of ν.

the parameters σ and ν, respectively. Further, Figure 2 indicates that the hrf of T has decreasing, unimodal and bimodal shapes.

Note that the parameters μ, σ and ν describe location, scale and skewness, for the failure times. For larger values of μ, survival times are larger and consequently the average of the failure time is larger. For larger values of σ, variability is larger and consequently the rate of acceleration (of the survival curves) is larger resulting in a higher hazard rate. Low values of ν indicating bimodality is more likely.

3 Regression models

In practical applications, the lifetimes of patients are affected by explanatory variables like age, tumor size, lymph node status and others. They can affect the probability of an individual being healed, so these variables need to be added in the statistical models to obtain better estimates as well as individual interpretations for such variables. Recently, a new cure rate survival regression model was proposed for predicting breast carcinoma survival in women who underwent mastectomy, modeling the probability of cure using explanatory variables.[18] Similarly, the generalized log-gamma regression model with cure fraction[8] was introduced to model the cured proportion with explanatory variables. The problem to model only the parameters relative to the cured proportion is that the explanatory variables also affect the lifetime of patients considered uncured, and therefore, it should be used to model the other parameters of the model. As an alternative to regression models cited above, the systematic part of the GAMLSS[19] can be expanded to allow not only the cure rate parameter but all parameters of the conditional distribution of T to be modeled as parametric functions of the explanatory variables.

3.1 Definition

Let $T \sim LSCp(t; \boldsymbol{\theta})$, where $\boldsymbol{\theta}^T = (\mu, \sigma, \nu, \tau)$ denotes the vector of parameters of the pdf ((equation (5)). Consider independent observations t_i's conditional on the parameter vector $\boldsymbol{\theta}_i$ (for $i = 1, 2, \ldots, n$) having pdf $f_p(t_i; \boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}^T = (\boldsymbol{\mu}^T, \boldsymbol{\sigma}^T, \boldsymbol{\nu}^T, \boldsymbol{\tau}^T)$ is a vector of parameters related to the response variable. We can define the elements of the vector $\boldsymbol{\theta}$ using four appropriate link functions as

$$\boldsymbol{\mu} = g_1(\mathbf{X}_1\boldsymbol{\beta}_1), \quad \boldsymbol{\sigma} = g_2(\mathbf{X}_2\boldsymbol{\beta}_2), \quad \boldsymbol{\nu} = g_3(\mathbf{X}_3\boldsymbol{\beta}_3), \quad \boldsymbol{\tau} = g_4(\mathbf{X}_4\boldsymbol{\beta}_4) \tag{9}$$

where $g_k(\cdot)$, for $k = 1, 2, 3, 4,$ denote the injective and twice continuously differentiable monotonic link functions, $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{m_k k})^T$ is a parameter vector of length $(m_k + 1)$, m_k denotes the number of explanatory variables related to the kth parameter and $\mathbf{X}_k$ is a known model matrix of order $n \times (m_k + 1)$. The total number of parameters to be estimated is given by $m = m_1 + m_2 + m_3 + m_4 + 4$ and the choice of parameters to be modeled by explanatory variables is discussed in Section 4. For the following sections, we shall consider the identity link function for $g_1(\cdot)$ and the logarithmic link function for $g_k(\cdot)$ ($k = 2$, 3, 4).

3.2 Inference

Consider a sample of n-independent observations $t_1, \ldots, t_n$. Let c_i denote the censoring time, $y_i = \min\{t_i, c_i\}$ and $\delta_i = I(t_i \leq c_i)$, where $\delta_i = 1$ if t_i is a time-to-event and $\delta_i = 0$ if it is right censored. From n observations, explanatory variables and censoring indicators $(y_1, \delta_1, \mathbf{x}_{k1}), \ldots, (y_n, \delta_n, \mathbf{x}_{kn})$, the log-likelihood function under non-informative censoring for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T, \boldsymbol{\beta}_4^T,)^T$ takes the form

$$\begin{aligned} l(\boldsymbol{\theta}) = & \sum_{i \in F} \left\{ \log(\tau_i) + \log(\nu_i) - \log(\sigma_i \pi) - \log(y_i) + \log\cosh(w_i) - \log\left[1 + \nu_i^2 \sinh^2(w_i)\right] \right\} \\ & - \sum_{i \in F} \sum_{i \in C} \tau_i \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu_i \sinh(w_i)] \right\} \end{aligned} \tag{10}$$

where $y_i = [\log(t_i) - \mu_i]/\sigma_i$, F and C denote the sets of individuals for which t_i is the log-lifetime or log-censoring and the vector of parameters are defined in equation (9) by specifying appropriate link functions for $g_k(\cdot)$, i.e., $\mu_i = \beta_{01} + \beta_{11}x_{i1} + \cdots + \beta_{m_k 1}x_{im_k}$.

The numerical maximization of the log-likelihood function (equation (10)) can be easily performed in the GAMLSS package in R. The advantage of this package is that we can use different maximization methods. Note that for censored observations, the additional package `gamlss.cens` is required to determine numerically the observed information of the likelihood function referring to the censored observations. The maximization algorithm adopted in the presence of censored data is the RS procedure.[12,19] This method is also available in the documentation of the GAMLSS package. For a specific data set, the likelihood potentially has multiple local maxima. This is investigated using different starting values and has generally not been found to be a problem in the data set analyzed, possibly due to the relatively large sample sizes used.

```
m1=gamlss(Surv(T,D)~x1+x2,
    sigma.formula =~x1+x2,
       nu.formula=~x1+x2,
      tau.formula=~x1+x2,
family=cens(‘LSCp’))
```

Here, we present an example of how to maximize the likelihood (equation (10)) in the R software. For the steps that will be presented below, consider the codes presented above. Let T be a response variable as well the failure indicator D. Now, consider the model `m1` where the explanatory variables X_1 and X_2 are used to model all parameters in equation (9). The results of the fitted model are accessed using `summary(m1)`. Note that for a null model (disregarding regression variables), the results obtained using this script still consider the regression structure ((equation (9)), e.g., $\tau = \exp(\beta_{04})$. The fit of the LSCp model gives the vector of estimated cured proportion

$$\widehat{\mathbf{p}} = \exp[-\exp(\mathbf{X}_4\widehat{\boldsymbol{\beta}}_4)], \quad 0 < \widehat{\mathbf{p}} < 1 \tag{11}$$

where $\mathbf{X}_4\widehat{\boldsymbol{\beta}}_4$ can be accessed using `m1$tau.fv`.

The asymptotic distribution of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is $N_m(\mathbf{0}, I(\boldsymbol{\theta})^{-1})$, where $I(\boldsymbol{\theta})$ is the expected information matrix. This asymptotic behavior holds if $I(\boldsymbol{\theta})$ is replaced by $\ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}})$, i.e., the observed information matrix evaluated at $\widehat{\boldsymbol{\theta}}$ given by $\ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|_{\hat{\theta}}$. The multivariate normal $N_m(\mathbf{0}, \ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}})^{-1})$ distribution can be used to construct approximate confidence intervals for the individual parameters.

Besides estimation of the model parameters, hypothesis tests can be investigated. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are disjoint subsets of $\boldsymbol{\theta}$. Consider the test of the null hypothesis $\mathcal{H}_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{01}$ against $\mathcal{H}_a : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{01}$, where $\boldsymbol{\theta}_{01}$ is a specified vector. Let $\widetilde{\boldsymbol{\theta}}$ be the restricted MLE of $\boldsymbol{\theta}$ obtained under $\mathcal{H}_0$. The likelihood ratio (LR) statistic to test $\mathcal{H}_0$ is given by $\Lambda = 2[\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widetilde{\boldsymbol{\theta}})]$. Under $\mathcal{H}_0$ and some regularity conditions, the LR statistic converges in distribution to a chi-square distribution with $\dim(\boldsymbol{\theta}_1)$ degrees of freedom.

An important consideration in the statistical analysis in the regression models is the assumption that all observations have equal variances. The non-compliance with this assumption affects the efficiency of the estimates of the parameters, so it is important to develop tests to determine the presence or absence of such homogeneity. Note that in healing models, there is heterogeneity in the data because of three subpopulations: one formed by the failure data, another for censored data and one formed by the cured individuals. In particular, we now consider the test for homogeneity of variance for the LSCp regression model with cure fraction based on the LR statistic. Following equations (5) and (6), we generalize the scale parameter σ by σ_i, where the parameter σ_i can be modeled by $\sigma_i = g_2(\mathbf{x}_{i2}^T\boldsymbol{\beta}_2)$, where $\mathbf{x}_{i2}$ is a vector of explanatory variable values. We assume that there exists a unique value σ_0, then $\sigma_i = \sigma_0$ and the Y_i's have constant variance. Hence, the LR statistic for the homogeneity of scalar parameter can be expressed by $\mathcal{H}_0 : \sigma_i = \sigma_0$ against $\mathcal{H}_a : \sigma_i \neq \sigma_0$, which is given by $\Lambda = 2[\ell(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\beta}}_4) - \ell(\widetilde{\boldsymbol{\beta}}_1, \sigma_0, \widetilde{\boldsymbol{\beta}}_3, \widetilde{\boldsymbol{\beta}}_4)]$, where $\widetilde{\boldsymbol{\beta}}_1$, $\widetilde{\boldsymbol{\beta}}_3$ and $\widetilde{\boldsymbol{\beta}}_4$ are the restricted MLEs of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$, respectively, obtained from the maximization of equation (10) under $\mathcal{H}_0 : \sigma_i = \sigma_0$. Analogously, we can perform the same tests of hypotheses for the parameters $\boldsymbol{\mu n}$, $\boldsymbol{v n}$ and $\boldsymbol{\tau n}$.

4 Model selection

Here, we consider the model selection process in four steps. The first step consists in choosing the best distribution to represent the lifetime and cure proportion. After, in the second step, we present a method to select the explanatory variables to fit each parameter of the selected model. The model assumptions are investigated in the third step. Finally, in the fourth step, we study the sensitivity of the chosen model with the existence of influential observations.

4.1 Select the distribution

```
AIC(m1)
BIC(m1)
deviance(m1)
```

In the first stage, the Akaike Information criterion (AIC), Bayesian Information criterion (BIC) and global deviance (GD) criteria are used to assess different fitted models. The GD, AIC and BIC criteria are defined by $GD = -2l(\widehat{\boldsymbol{\theta}})$, $AIC = GD + 2k$ and $BIC = GD + \log(n)k$, respectively, where $l(\widehat{\boldsymbol{\theta}})$ is the total log-likelihood function, n represents the sample size and k denotes the number of fitted parameters. The model with the smallest values for these criteria is then selected. The codes to access these statistics are presented above.

4.2 Selecting explanatory variables

For the LSCp GAMLSS regression, the selection of the terms for all parameters is performed using a stepwise AIC procedure.[20] There are many different strategies that could be applied for selection of the terms used to model the four parameters $\boldsymbol{\mu n}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu n}$ and $\boldsymbol{\tau n}$. 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 distribution and link functions, the strategy consists of two steps. In the first step, we adopt a forward selection procedure to select an appropriate model for $\boldsymbol{\mu n}$, with $\boldsymbol{\sigma n}$, $\boldsymbol{\nu n}$ and $\boldsymbol{\tau n}$ fitted as constants. After that, repeat the same procedure to select the model for $\boldsymbol{\sigma n}$, $\boldsymbol{\nu n}$ and $\boldsymbol{\tau n}$, 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 $\boldsymbol{\nu n}$, with $\boldsymbol{\mu n}$, $\boldsymbol{\sigma n}$ and $\boldsymbol{\tau n}$ fitted as constants and repeat this procedure for $\boldsymbol{\sigma n}$ and $\boldsymbol{\mu n}$, respectively. At the end of the steps described above, the final model may contain different subsets from χ for $\boldsymbol{\mu n}$, $\boldsymbol{\sigma n}$, $\boldsymbol{\nu n}$ and $\boldsymbol{\tau n}$.

```
m1=gamlss(Surv(T,D)~1,family=cens(‘LSCp’))
m2=stepGAICAll.A(m1,scope=list(lower=~1,
                          upper=~x1+x2+x3))
```

An easy way to reproduce the steps mentioned above is using the `stepGAICAll.A` function implemented in `GAMLSS` package. The first step consists of fitting a null model `m1` (without regression structure) considering the lifetime T variable as well as the failure indicator D. Next, consider the second model `m2`, in which all parameters can be modeled by the explanatory variables indicated in the `upper` command. An example is shown in the codes presented above, which has three explanatory variables, X_1, X_2 and X_3. At the end, the final model `m2` may contain different subsets from χ for μ, σ, ν and τ.

4.3 Diagnostics

In order to study departures from the error assumption and the presence of outlying observations, we can use the diagnostic tools in the GAMLSS package. The first technique consists of the normalized randomized quantile residuals,[21] which are given by $\hat{r}_i = \Phi^{-1}(\hat{u}_i)$, where $\Phi^{-1}(\cdot)$ is the quantile function (qf) of the standard normal variate and $\hat{u}_i = F(t_i|\widehat{\boldsymbol{\theta}}_i)$. For censored response variables, $\hat{u}$ is defined as a random value from a uniform distribution on the interval $[1 - S(t_i|\widehat{\boldsymbol{\theta}}_i), 1]$.

```
plot(density(m2$residuals))
qqnorm(m2$residuals)
qqline(m2$residuals,col=2)
wp(m2)
```

Although the quantile residuals are widely used in literature, it is not possible to identify specifically failures to fit the mean, variance, skewness and kurtosis existing in the variable responses. As an alternative, we can use the Worm Plots (WP).[22] These plots of the residuals were introduced in order to identify regions (intervals) of an explanatory variable within which the model does not fit adequately the data. This is a diagnostic tool for checking the residuals for different ranges of one or two explanatory variables. The idea consists to fit cubic models to each of the detrended QQ plots with the resulting constant, linear, quadratic and cubic coefficients, thus indicating differences between the empirical and model residual mean, variance, skewness and kurtosis, respectively, within the range in the QQ plot. The interpretations of the shapes of the WP are: a vertical shift, a slope, a parabola or a S shape, thus indicating a misfit in the mean, variance, skewness and excess kurtosis of the residuals, respectively. Let `m2` the final model selected. Using the commands presented in the box, we can easily access the residuals discussed before.

4.4 Global influence

Since regression models are sensitive to the underlying model assumptions, performing a sensitivity analysis is strongly advisable. This idea was used to motivate the assessment of influence analysis,[23] suggesting that more confidence can be put in a model, which is relatively stable under small modifications. The best known perturbation schemes are based on case-deletion,[24] in which the effects or perturbations of completely removing cases from the analysis are studied.

In the following, a quantity with subscript "$(-i)$" refers to the original quantity with the i th case deleted. For model (9), the log-likelihood function ((equation (10)) for $\boldsymbol{\theta}$ is denoted by $l(\boldsymbol{\theta})$. Let $\widehat{\boldsymbol{\theta}}^T_{(-i)} = \left(\widehat{\boldsymbol{\mu}}^T_{(-i)}, \widehat{\boldsymbol{\sigma}}^T_{(-i)}, \widehat{\boldsymbol{\nu}}^T_{(-i)}, \widehat{\boldsymbol{\tau}}^T_{(-i)}\right)$ be the MLEs of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$ obtained from $l(\boldsymbol{\theta}_{(-i)})$. To assess the influence of the i th case on the MLE $\widehat{\boldsymbol{\theta}}$, the idea is to compare the difference between $\widehat{\boldsymbol{\theta}}_{(-i)}$ and $\widehat{\boldsymbol{\theta}}$. If deletion of a case seriously influences the estimates, more attention should be given to that case. Hence, if $\widehat{\boldsymbol{\theta}}_{(-i)}$ is far from $\widehat{\boldsymbol{\theta}}$, then the i th case is regarded as an influential observation. A popular measure of the difference between $\widehat{\boldsymbol{\theta}}_{(-i)}$ and $\widehat{\boldsymbol{\theta}}$, called log-likelihood distance, is given by

$$LD_i(\boldsymbol{\theta}) = 2\left[l(\widehat{\boldsymbol{\theta}}) - l\left(\widehat{\boldsymbol{\theta}}_{(-i)}\right)\right]$$

Note that for the GAMLSS, all parameters can be modeled by explanatory variables, so the log-likelihood can potentially have multiple local maxima. We suggest to use the MLE $\widehat{\boldsymbol{\theta}}$ as initial vector to obtain the MLE $\widehat{\boldsymbol{\theta}}_{(-i)}$. An example of how to calculate $LD_i(\boldsymbol{\theta})$ using the `GAMLSS` package is given in supplemental material.

5 Simulation study

In this section, we report a Monte Carlo simulation study assessing the finite sample behavior of the MLEs of the parameters for different sample sizes, cured percentages and percentage of censored in the failure times. Note that cured percentages represent the percentage of individuals who are considered cured and the censored failure time percentages represent the percentages of individuals who for some reason did not remain until the end of the study. The cured percentage is denoted by p as shown in equation (11) and the censored failure times percentage is denoted by ψ.

We can simulate LSCp random variables using the qf, which is obtained by inverting $F(t) = 1 - S(t) = u$, where $S(t)$ represents the survival function for non-censored observations (equation (7)). The qf of $T \sim \text{LSCp}(t, \mu, \sigma, \nu, \tau)$ is given by

$$T = Q(u) = \exp\left(\mu + \sigma \text{ arcsinh}\left\{\frac{1}{\nu}\tan[\pi(k(u,\tau) - 0.5)]\right\}\right) \tag{12}$$

where $k(u,\tau) = -\log[(u-1)(\mathrm{e}^{-\tau} - 1)]$. Equation (12) can be used for simulating random variables by fixing μ, σ, ν, τ and setting u as a uniform random variable in the $(0,1)$ interval.

```
rLSCp(n,mu,sigma,nu,tau)
```

To generate the cured proportion, we adopt the following strategy. Let n be the total sample size, composed by the sample of the cured individuals C, with size $n_c = n\mathrm{e}^{-\tau}$, and by the sample of the observed times T, with size $n_t = n - n_c$. Now, we generate n_t observations using (12) and, for generate n_c cured observations, we consider that

$C \sim U[\max(T), 2 \times sd(T)]$, where $sd(T)$ represents the standard deviation of the generated time sample. The samples can be easily generated in R using the codes presented above. Censored failure times can be set by selecting random values in T generated samples.

Here, we consider that the lifetimes T are composed by the lifetimes of two groups, g_1 and g_2, where $T|g_1 \sim LSCp(\mu_1 = 1.5, \sigma_1 = 0.3, \nu_1 = 0.1, \tau_1 = 2)$ and $T|g_2 \sim LSCp(\mu_2 = 2.5, \sigma_2 = 0.2, \nu_2 = 0.5, \tau_2 = 1)$. For each group, samples of size $n_g = 25$, 50 and 75 are generated for each replication, yielding the total sample sizes $n = 50$, 100 and 150. The cured percentage for g_1 and g_2 are $p_1 = 0.135$ and $p_2 = 0.367$, respectively. We also consider different censored failure time percentages, $\psi = 0, 0.1$, where the number of censored failure time for g_1 and g_2 are given by $n_g(1 - p_1)\psi$ and $n_g(1 - p_2)\psi$, respectively. For $\psi = 0.1$, the total censoring percentages for g_1 and g_2 are 22.1% and 43.1%, respectively. The codes used in this section are presented in supplemental material.

Using equation (9), we can define the regression structure as

$$\mu_i = \beta_{01} + \beta_{11}x_{1i}, \quad \sigma_i = \exp(\beta_{02} + \beta_{12}x_{1i}), \quad \nu_i = \exp(\beta_{03} + \beta_{13}x_{1i}), \quad \tau_i = \exp(\beta_{04} + \beta_{14}x_{1i})$$

where $x_{1i} = 1$ and $x_{1i} = 0$ represent the groups g_1 and g_2, respectively. The model parameters are defined by $\mu_1 = \beta_{01} + \beta_{11}$, $\mu_2 = \beta_{01}$, $\sigma_1 = \exp(\beta_{02} + \beta_{12})$, $\sigma_2 = \exp(\beta_{02})$, $\nu_1 = \exp(\beta_{03} + \beta_{13})$, $\nu_2 = \exp(\beta_{03})$, $\tau_1 = \exp(\beta_{04} + \beta_{14})$ and $\tau_2 = \exp(\beta_{04})$.

The lifetimes considered in each fit are evaluated as $\min(t_i, c_i)$ and, for each configuration of n and ψ, all results are obtained from 1000 Monte Carlo replications. For each replication, we evaluate the MLEs 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 codes presented above are used for maximizing the total log-likelihood function ((equation (10)).

Table 1. The AEs, biases and MSEs based on 1000 simulations for the LSCp model when $\mu_1 = 1.5$, $\sigma_1 = 0.3$, $\nu_1 = 0.1$, $\tau_1 = 2$, $\mu_2 = 2.5$, $\sigma_2 = 0.2$, $\nu_2 = 0.5$ and $\tau_2 = 1$.

ψ	n	θ	AE	Bias	MSE	θ	AE	Bias	MSE
0%	50	μ_1	1.540	0.040	0.028	μ_2	2.592	0.092	0.055
		σ_1	0.290	0.010	0.005	σ_2	0.194	0.006	0.007
		ν_1	0.101	0.001	0.014	ν_2	0.412	0.088	0.181
		τ_1	2.198	0.198	0.095	τ_2	1.162	0.162	0.100
0%	100	μ_1	1.514	0.014	0.013	μ_2	2.527	0.027	0.013
		σ_1	0.297	0.003	0.002	σ_2	0.198	0.002	0.003
		ν_1	0.101	0.001	0.004	ν_2	0.490	0.010	0.085
		τ_1	2.028	0.028	0.041	τ_2	1.058	0.058	0.016
0%	150	μ_1	1.508	0.008	0.006	μ_2	2.505	0.005	0.005
		σ_1	0.296	0.004	0.002	σ_2	0.200	0.000	0.002
		ν_1	0.098	0.002	0.002	ν_2	0.507	0.007	0.052
		τ_1	2.042	0.042	0.019	τ_2	1.001	0.001	0.003
10%	50	μ_1	1.536	0.036	0.034	μ_2	2.637	0.137	0.079
		σ_1	0.288	0.012	0.005	σ_2	0.192	0.008	0.007
		ν_1	0.096	0.004	0.009	ν_2	0.361	0.139	0.139
		τ_1	2.004	0.004	0.112	τ_2	1.069	0.069	0.109
10%	100	μ_1	1.516	0.016	0.013	μ_2	2.530	0.030	0.023
		σ_1	0.293	0.007	0.002	σ_2	0.197	0.003	0.004
		ν_1	0.097	0.003	0.003	ν_2	0.482	0.018	0.103
		τ_1	1.835	0.165	0.035	τ_2	0.967	0.033	0.021
10%	150	μ_1	1.509	0.009	0.006	μ_2	2.510	0.010	0.009
		σ_1	0.294	0.006	0.002	σ_2	0.199	0.001	0.003
		ν_1	0.096	0.004	0.002	ν_2	0.507	0.007	0.072
		τ_1	1.854	0.146	0.016	τ_2	0.897	0.103	0.006

AE: average estimates; MSE: mean squared error.

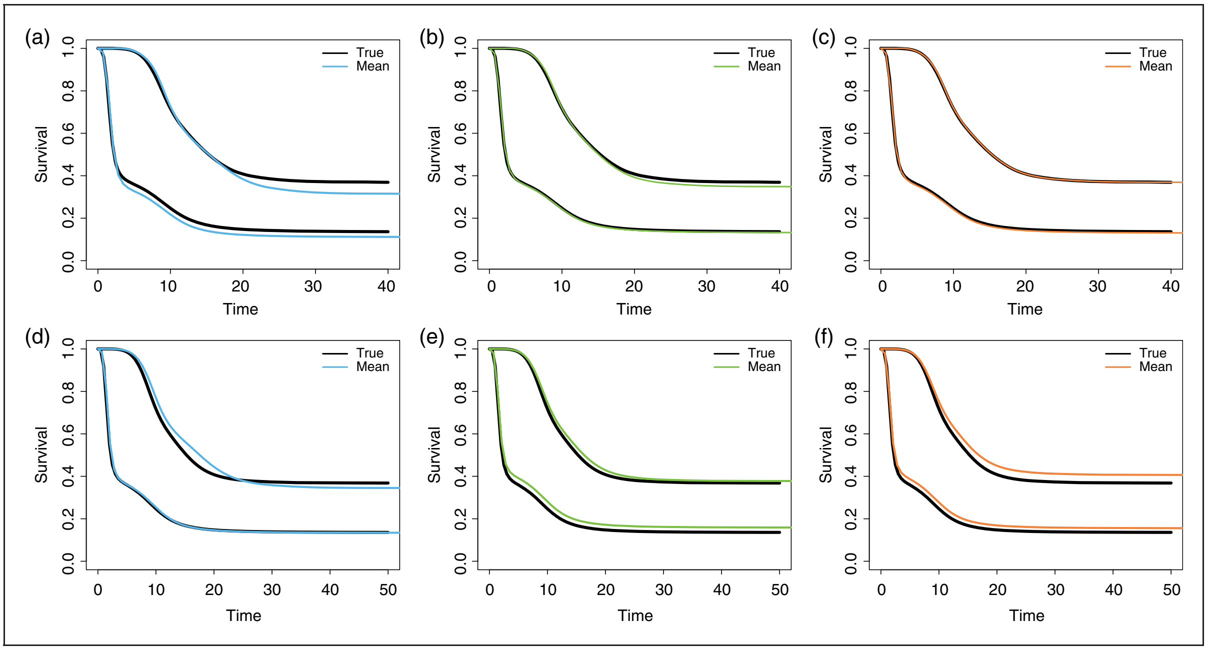

Figure 3. LSCp survival functions at the true parameter values and at the AEs obtained in Table 1 by taking $\psi = 0$ (a) $n = 50$, (b) $n = 100$ and (c)$n = 150$ and by taking $\psi = 0.1$ (d) $n = 50$, (e) $n = 100$ and (f)$n = 150$.

The results are reported in Table 1 and, for a visual analysis, we present in Figure 3 the generated and the estimated (considering the AEs given in Table 1) survival functions for $n = 50$, 100 and 150 and considering the two groups represented by the explanatory variable x_{1i}.

The results of the Monte Carlo study in Table 1 indicate that the MSEs of the MLEs of the parameters 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 MLEs. The normal approximation can often be improved by using bias adjustments to these estimators. In general, for the LSCp GAMLSS, the variances and MSEs increase when the failure times percentage ψ increases, as expected. Even with high percentages of censored observations, we can note a good fit of the LSCp GAMLSS. This fact can be noted in Figure 3.

6 Predicting breast cancer data

The highest breast cancer incidence rates continue to be observed in high-income countries, including countries in Northern America, Australia, and Northern and Western Europe. Almost 1.7 million new breast cancer cases and 521,900 breast cancer deaths were estimated to have occurred in 2012 worldwide.[25] One in eight women (12%) are expected to have this diagnosis in her lifetime. Although breast cancer incidence rates continued to increase in many countries, mortality rates have declined in 34 of 57 countries. These reductions have been attributed to early detection through mammography and improved treatment.

The initial prognostic model considers the explanatory variables tumor size, histology grade and lymph node status as basic factors to be taken into consideration.[26] Due the fact of the introduction of new imaging modalities, the multifocality has also been considered as a important prognostic to be taken into consideration. The results using magnetic resonance imaging reveal that the multifocality appears in a considerable proportion of cases, thus influencing some clinicians to take this information into account when planning surgical and oncologic therapy.[27] Surgery is the most common treatment for breast cancer. There are several kinds of surgery. The surgeon usually removes one or more lymph nodes from under the arm to check for cancer cells. If cancer cells are found in the lymph nodes, other cancer treatments will be needed. At any stage of disease, care is available to control pain and other symptoms to relieve the side effects of treatment, and to ease emotional concerns.

The data set represents the survival times (T) until the patient's death or the censoring times at the end of the study.[28] A total of $n = 284$ women who had been treated with mastectomy and axillary lymph node dissection at Memorial Sloan-Kettering Cancer Center (New York, NY) between 1976 and 1979 met the following

Table 2. MLEs of the LSCp model parameters, the corresponding SEs (given in parentheses) and the GD, AIC and BIC statistics.

μ	e^{σ}	e^{ν}	e^{τ}	GD	AIC	BIC
2.271 (0.057)	−0.987 (0.055)	−0.960 (0.096)	−0.853 (0.060)	712.8	720.8	735.4

GD: global deviance; AIC: Akaike Information criterion; BIC: Bayesian Information criterion.

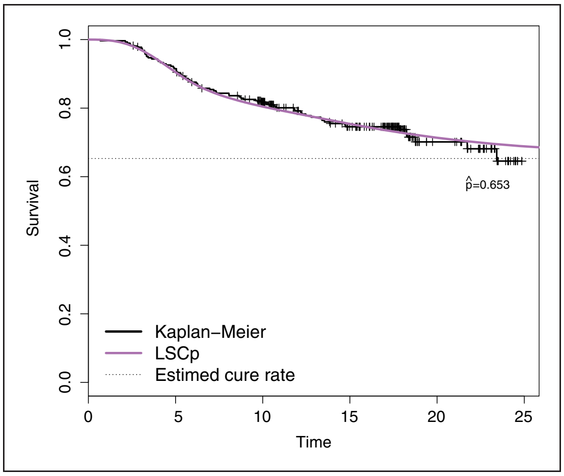

Figure 4. The estimated and empirical survival functions.

requirements for study inclusion: confirmation of the presence of invasive mammary carcinoma, no receipt of neoadjuvant or adjuvant systemic therapy, no previous history of malignancy, and negative lymph node status as assessed on routine histopathologic examination. There are 74% censored observations corresponding to the women who died from other causes or were still alive at the end of the study.

Some explanatory variables are associated with pathologic characteristics of the tumor. The tumor grading was performed using the standard modified Bloom–Richardson system. The lymphovascular invasion was obtained using morphologic criteria. The lymph node status was measured according to immunohistochemistry (IHC) and hematoxylin and eosin (H&E) stains. The explanatory variables for each woman ($i = 1, \ldots, 284$) are described below:

- t_i: observed time (in years);
- δ_i: failure indicator (0: censored, 1: observed);
- x_{i1}: age (in years);
- x_{i2}: multifocality (0: no, 1:yes);
- x_{i3}: tumor size (in cm);
- x_{i4}: tumor grading (0: I, 1: II, III and lobular);
- x_{i5}: lymphovascular invasion (0: no, 1: yes)
- x_{i6}: lymph node status (0: IHC+ IHC- and H&E-, 1: IHC+ and H&E+).

We start the analysis by fitting the LSCp model (9) disregarding regression variables. Table 2 gives the MLEs (and the corresponding SEs in parentheses) of the model parameters and the values of the GD, AIC and BIC statistics for the fitted model. Using equation (11), the estimated cure proportion is given by $\hat{p} = \exp(-0.853) = 0.653$, being an indication of the presence of a proportion of patients for whom the breast carcinoma will never recur.[4] Then, the patients can be considered as cured. Figure 4 provides the plots of the estimated and empirical survival function. Table 2 and Figure 4 indicate that the LSCp model provides a good fit to these data.

Recently, the Poisson beta Weibull (PBW), Poisson Weibull (PW), negative binomial beta Weibull (NBiBW), negative binomial Weibull (NBiW), geometric beta Weibull (GBW) and geometric Weibull (GW) cure rate

Table 3. The GD, AIC and BIC statistics for some models.

Fitted models	GD	AIC	BIC
LSCp	**670.3**	**690.3**	**726.8**
PBW	674.2	696.2	736.3
PW	678.9	696.9	729.7
NBiBW	673.1	697.1	740.8
NBiW	678.9	698.9	735.3
GBW	675.5	697.5	737.6
GW	680.2	698.2	731.0

GD: global deviance; AIC: Akaike Information criterion; BIC: Bayesian Information criterion; LSCp: log-sinh Cauchy promotion time; PBW: Poisson beta Weibull; PW: Poisson Weibull; NBiBW: negative binomial beta Weibull; NBiW: negative binomial Weibull; GBW: geometric beta Weibull; GW: geometric Weibull. Bold figures highlight the lowest value.

regression models were fitted to these data[18] using all the explanatory variables to model the cured proportion parameter. We compare the results of these models by fitting the LSCp regression model, in which all explanatory variables are used to model τ, i.e.

$$\log \boldsymbol{\tau} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{X}_1 + \boldsymbol{\beta}_2\mathbf{X}_2 + \boldsymbol{\beta}_3\mathbf{X}_3 + \boldsymbol{\beta}_4\mathbf{X}_4 + \boldsymbol{\beta}_5\mathbf{X}_5 + \boldsymbol{\beta}_6\mathbf{X}_6$$

The values of the GD, AIC and BIC statistics for the fitted models are listed in Table 3. The lowest values of the information criteria correspond to the LSCp model, which provides a better fit to the current breast cancer data than the other models.

Using the steps described in Section 4 to select the additive terms for the different parameters, we present results for the model parameters defined by

$$\mu_i = \beta_{01} + \beta_{41}x_{i4},\ \sigma_i = \exp(\beta_{02} + \beta_{22}x_{i2} + \beta_{62}x_{i6}),$$
$$\nu_i = \exp(\beta_{03} + \beta_{53}x_{i5}) \text{ and } \tau_i = \exp(\beta_{04} + \beta_{34}x_{i3} + \beta_{44}x_{i4} + \beta_{64}x_{i6})$$

As suggested by a referee, we compare the results by fitting the Weibull cure rate mixture (Weibullcr) model with scale $\mu > 0$, shape $\sigma > 0$ and cure rate $\nu \in [0, 1]$ parameters. The Weibullcr model was also implemented in the GAMLSS package, which the codes can be found in the supplemental material for future research. The additive terms selected for the Weibullcr model are

$$\mu_i = \exp(\beta_{01} + \beta_{41}x_{i4} + \beta_{51}x_{i5}),\ \sigma_i = \exp(\beta_{02})$$

and

$$\nu_i = logit(\beta_{03} + \beta_{23}x_{i2} + +\beta_{33}x_{i3} + \beta_{43}x_{i4} + \beta_{53}x_{i5} + \beta_{63}x_{i6})$$

Table 4 provides the MLEs, SEs and p values obtained from the fitted LSCp and Weibullcr GAMLSS regressions. We note that all parameters are significant at the 5% significance level, indicating the accuracy of the method to select the additive terms. Based on the figures in this table, we can conclude that the explanatory variables tumor size, tumor grading and lymph node status are significant factors for the cure probability of women with breast cancer. The variables tumor grading and lymph node status are also significant to model the location and scale parameters. It means that these variables have influence in the mean and variance in the women's lifetimes who were considered uncured. Finally, the variables multifocality and lymphovascular invasion are significant to model the variability and symmetry existing in the lifetime of the uncured women. Note that the parameter estimates, relative to the cure parameter, from LSCp GAMLSS "τ" are different to the parameter estimates from Weibullcr GAMLSS "ν." This happens because the link functions are not the same. Moreover, the SEs of the MLEs from the fitted LSCp GAMLSS are smaller than those obtained from the Weibullcr GAMLSS. This fact indicates that the estimates of the LSCp model are more precise than those of the Weibullcr GAMLSS. A difference exists regarding the significance of the covariate X_2 and X_5, because they are non-significant in the LSCp model, whereas they become significant at the 5% level in the Weibullcr GAMLSS.

Table 4. The MLEs, corresponding SEs and p values of the estimates from the fitted LSCp and Weibullcr GAMLSS regression.

Model	Parameter	Estimate	SE	p	Parameter	Estimate	SE	p
LSCp	β_{01}	1.550	0.052	<0.001	β_{53}	1.202	0.205	<0.001
	β_{41}	0.692	0.064	<0.001	β_{04}	−4.400	0.187	<0.001
	β_{02}	−1.016	0.043	<0.001	β_{34}	0.288	0.060	<0.001
	β_{22}	−0.464	0.101	<0.001	β_{44}	1.205	0.197	<0.001
	β_{62}	−0.625	0.074	<0.001	β_{64}	2.932	0.174	<0.001
	β_{03}	−1.511	0.097	<0.001				
Weibullcr	β_{01}	0.711	0.106	<0.001	β_{23}	−1.358	0.438	0.002
	β_{41}	1.602	0.113	<0.001	β_{33}	−0.647	0.109	<0.001
	β_{51}	0.806	0.108	<0.001	β_{43}	−6.030	0.218	<0.001
	β_{02}	0.410	0.043	<0.001	β_{53}	−4.061	0.468	<0.001
	β_{03}	8.562	0.243	<0.001	β_{63}	−2.816	0.411	<0.001

SE: standard error.

Table 5. LR tests.

Parameter	$l(\boldsymbol{\theta})$	Λ	p	Parameter	$l(\boldsymbol{\theta})$	Λ	p
Complete	−327.689	–	–	β_{53}	−330.979	6.581	0.010
β_{41}	−329.674	3.970	0.046	β_{34}	−331.613	7.849	0.005
β_{22}	−330.143	4.909	0.027	β_{44}	−334.250	13.123	0.001
β_{62}	−332.200	9.022	0.003	β_{64}	−333.817	12.257	0.001

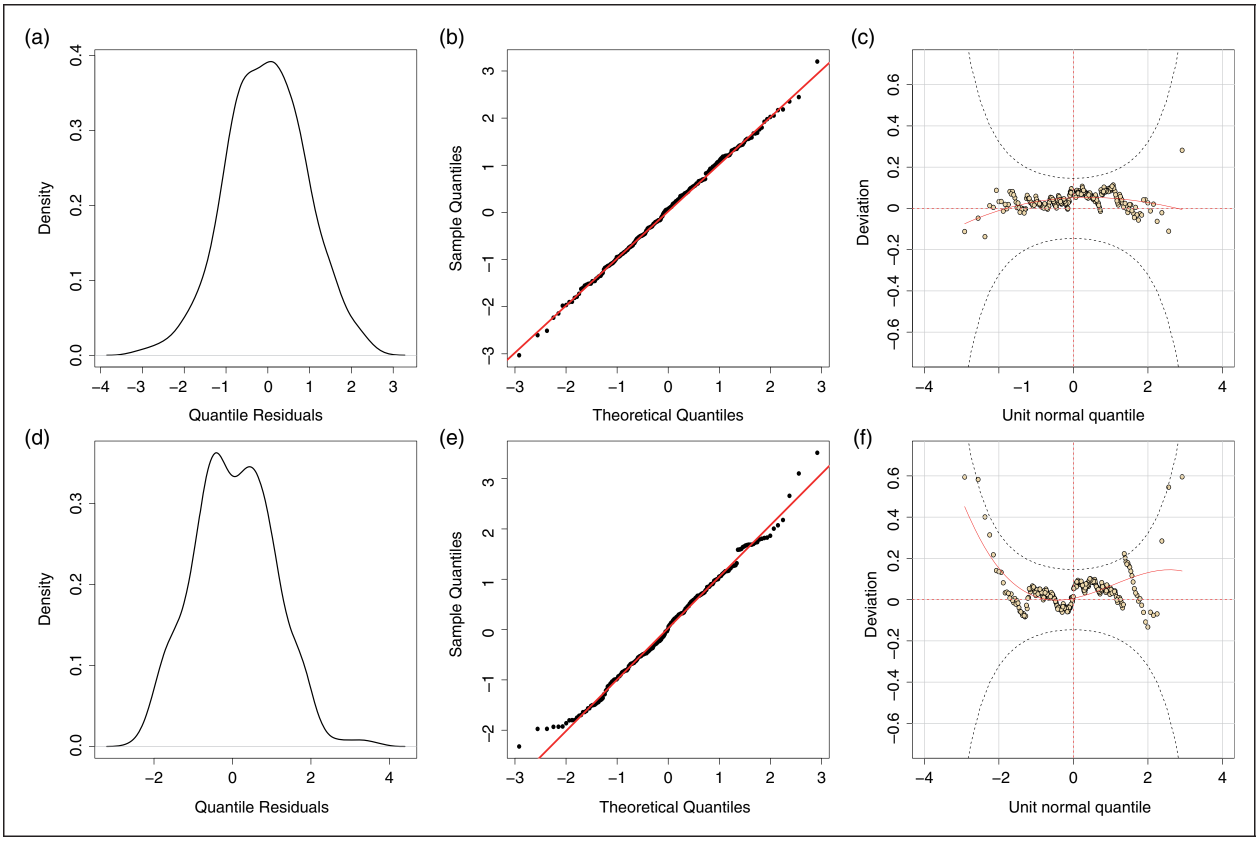

Figure 5. Residual analysis: For the LSCp and Weibullcr models, (a)-(d) Density of the quantile residuals, (b)-(e)Q-Q plot and (c)-(f) WP, respectively.

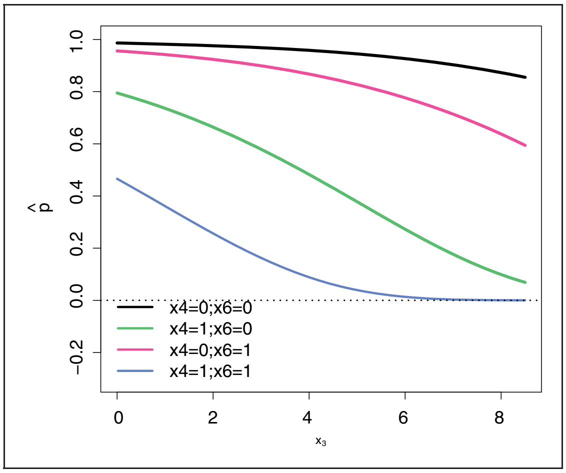

Figure 6. The estimated cured proportions for each level of X_4 and X_6 for all range of X_3.

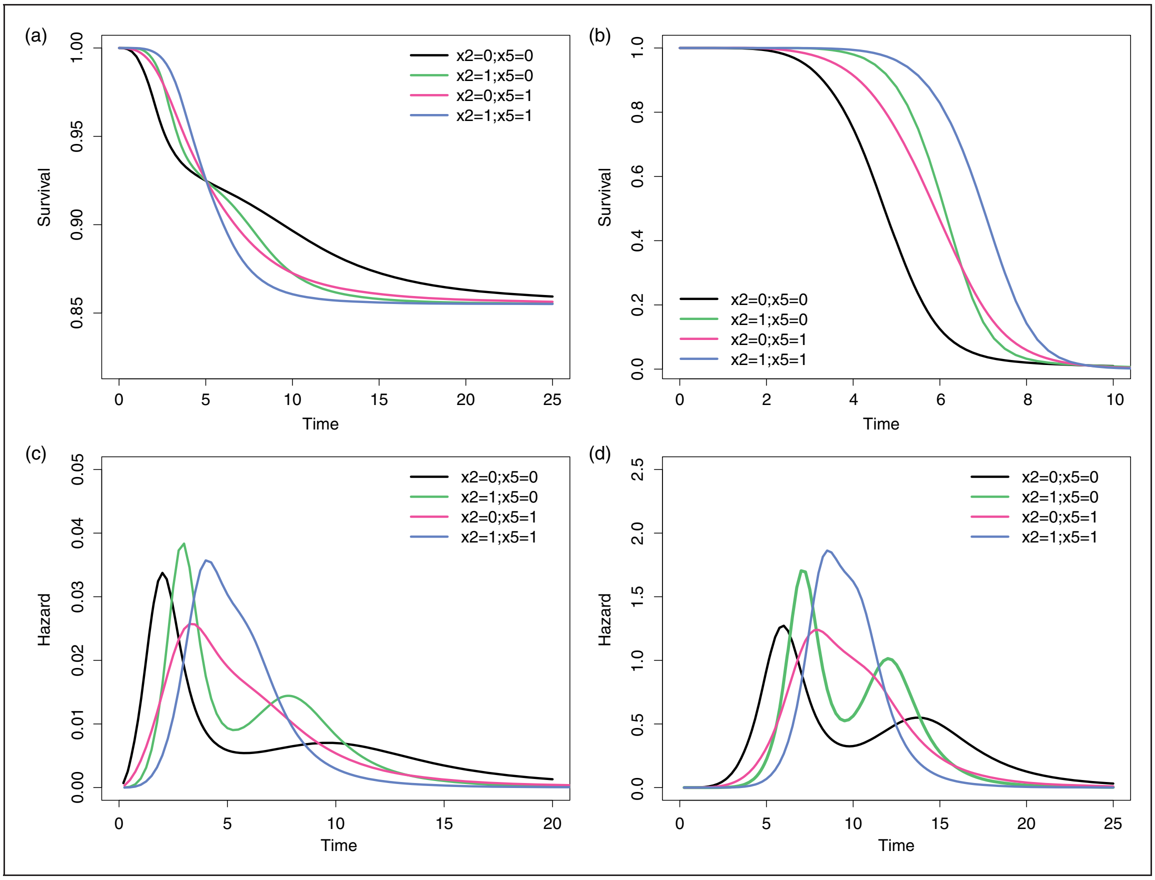

Figure 7. For maximum tumor size "max(X_3)", the estimated survival functions for (a) g_2 and $(b)\, g_1$ as well as the fitted hazard functions for (c) g_2 and $(d)\, g_1$.

Table 5 provides the formal tests to verify the significance of the explanatory variables presented in Table 4 for the LSCp model. Using the LR test, we compare the complete model with submodels, removing each explanatory variable selected. For example, to test if the explanatory variable x_{i2} indeed need to be used to model the scale parameter, we can test the hypothesis $\mathcal{H}_0 : \beta_{22} = 0$. We can conclude, at the 5% significance level, that all selected explanatory variables should remain in the selected model.

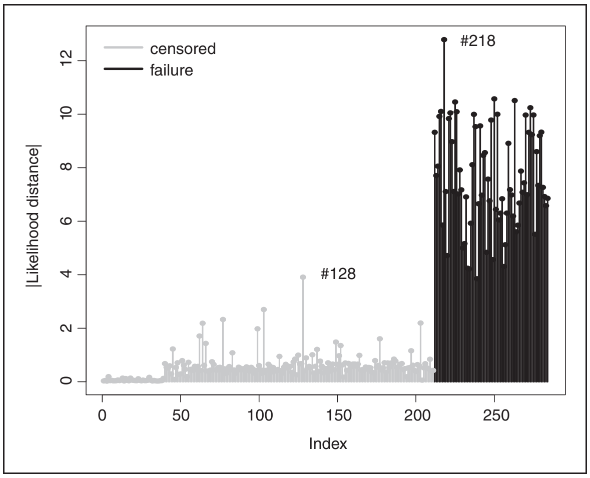

Figure 8. Index plots for $|LD_i(\boldsymbol{\theta})|$.

The criteria obtained for the fitted models in Table 4 are GD = 655.3, AIC = 677.3 and BIC = 717.5 for the fitted LSCp GAMLSS and GD = 661.2, AIC = 681.2 and BIC = 717.7 for the fitted Weibullcr GAMLSS. The plots of residual analysis are displayed in Figure 5 in order to verify the adequacy and the assumptions of the fitted models. In Figure 5(a) and (b), we note that the quantile residuals have an approximately normal distribution. The WP given in Figure 5(c) reveals that the proposed regressions for modeling the mean, variance, skewness and kurtosis are correct. Figure 5(d) and (e) indicates that the Weibullcr model does not present a good fit for extreme values. Also, in Figure 5(f), we can note a U-shape in the WP, thus indicating failure for modeling the skewness in the data. We can conclude from this plot that the proposed model provides a good fit for the breast cancer data.

Using equation (11), the estimated cured proportions can be determined using the results obtained in equation (4) as $p_i = \exp[-\exp(-4.290 + 2.817x_{i4} + 1.195x_{i6} + 0.288x_{i3})]$. In Figure 6, we present the estimated cured proportions for different levels of the explanatory variables X_4 and X_6 as functions of X_3. We note in this plot that the tumor grading II, III and lobular are very aggressive, influencing dramatically the cured probability. It is also possible to note that the tumor size has a large influence on the probability of cure in patients with tumors classified as II, III and lobular with lymph node status IHC+ and H&E+.

We define the high-risk g_1 group composed by $X_4 = 1$ and $X_6 = 1$ (blue line in Figure 6) and the low-risk g_2 group composed by $X_4 = 0$ and $X_6 = 0$ (black line in Figure 6). In Figure 7, we present the fitted survival functions for g_1 and g_2 considering the maximum of tumor size $\max(X_3) = 8.5$. We also present in this plot the fitted hazard functions for g_1 and g_2. We can observe in these plots the effects of X_2 and X_5 in the scale and symmetry parameters, respectively.

Next, we compute the case deletion measures $LD_i(\boldsymbol{\theta})$. Figure 8 displays the plots of the absolute influence measure index. We note that the cases #128 and #218 are possible influential observations. The censored observation #128 has a highest tumor size X_3 and #128 corresponds to the highest lifetime $t_i = 18.75$ for the g_1 group when $X_2 = 0$ and $X_5 = 1$ (see Figure 7(b) pink curve).

7 Conclusions

The parametric log-sinh Cauchy promotion time generalized additive model for location, scale and shape (LSCp GAMLSS) regression provides a flexible model for a dependent real outcome. The parameters of the model can be interpreted as relating to location, scale, skewness/bimodality and cure rate, and they can each be modeled as parametric functions of explanatory variables. Procedures for fitting the LSCp GAMLSS regression and for model diagnostics are included in the GAMLSS package, which are available from the authors. We use the proposed model to estimate breast carcinoma mortality, assuming that the number of competing causes that can influence the survival time follows a Poisson distribution. The results reveal that the tumor size, tumor grading and lymph node status have a significant influence in the cure probability. We also conclude that the variables tumor grading, lymph node status, multifocality and lymphovascular invasion are also significant to model the women's lifetimes who were considered uncured.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first author acknowledge the financial support of the "Ciência sem Fronteiras" program of CNPq (Brazil) under the process number 200574/2015-9.

Supplemental material

Supplemental material is available for this article online.

References

1. Berkson J and Gage RP. Survival curve for cancer patients following treatment. *J Am Stat Assoc* 1952; **47**: 501–515.
2. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J R Stat Soc B* 1949; **11**: 15–53.
3. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982; **38**: 1041–1046.
4. Yakovlev A and Tsodikov AD. Stochastic models of tumor latency and their biostatistical applications. In: *Mathematical biology and medicine*. Vol. 1, Hackensack, NJ: World Scientific, 1996, pp.321–326.
5. de Castro M, Cancho VG and Rodrigues J. A hands-on approach for fitting long-term survival models under the GAMLSS framework. *Comput Meth Progr Biomed* 2010; **97**: 168–177.
6. Hellwig B, Hengstler JG, Schmidt M, et al. Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformat* 2010; **11**: 1.
7. Ramires TG, Ortega EMM, Cordeiro GM, et al. A bimodal flexible distribution for lifetime data. *J Stat Comput Simulat*, Epub ahead of print 8 June 2015. DOI:10.1080/00949655.2015.1115047
8. Ortega EM, Cancho VG and Paula GA. Generalized log-gamma regression models with cure fraction. *Lifetime Data Anal* 2009; **15**: 79–106.
9. Silva GO, Ortega EMM, Cancho VG, et al. Log-Burr XII regression models with censored data. *Comput Stat Data Analy* 2008; **52**: 3820–3842.
10. Hashimoto EM, Ortega EMM, Cordeiro GM, et al. The Log-Burr XII regression model for grouped survival data. *J Biopharmaceut Stat* 2012; **22**: 141–159.
11. Rodrigues J, de Castro M, Cancho VG, et al. COM—Poisson cure rate survival models and an application to a cutaneous melanoma data. *J Stat Plan Infer* 2009; **139**: 3605–3611.
12. Stasinopoulos DM and Rigby RA. Generalized additive models for location scale and shape (GAMLSS) in R. *J Stat Softw* 2007; **23**: 1–46.
13. R Core Team. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, www.R-project.org/ (2015, accessed 8 February 2017).
14. Calsavara VF, Tomazella VL and Fogo JC. The effect of frailty term in the standard mixture model. *Chil J Stat* 2013; **4**: 95–109.
15. Cooner F, Banerjee S, Carlin BP, et al. Flexible cure rate modeling under latent activation schemes. *J Am Stat Assoc* 2007; **102**: 560–572.
16. Ibrahim JG, Chen MH and Sinha D. *Bayesian survival analysis*. New York: Springer, 2001.
17. Li CS, Taylor JM and Sy JP. Identifiability of cure models. *Stat Prob Lett* 2001; **54**: 389–395.
18. Ortega EM, Cordeiro GM, Campelo AK, et al. A power series beta Weibull regression model for predicting breast carcinoma. *Stat Med* 2015; **34**: 1366–1388.
19. Rigby RA and Stasinopoulos DM. Generalized additive models for location, scale and shape. *J R Stat Soc C* 2005; **54**: 507–554.
20. Voudouris V, Gilchrist R, Rigby R, et al. Modelling skewness and kurtosis with the BCPE density in GAMLSS. *J Appl Stat* 2012; **39**: 1279–1293.
21. Dunn PK and Smyth GK. Randomized quantile residuals. *J Comput Graph Stat* 1996; **5**: 236–244.
22. Buuren SV and Fredriks M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Stat Med* 2001; **20**: 1259–1277.
23. Cook RD. Assessment of local influence. *J R Stat Soc B* 1986; **48**: 133–169.
24. Cook RD and Weisberg S. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.
25. DeSantis CE, Bray F, Ferlay J, et al. International variation in female breast cancer incidence and mortality rates. *Cancer Epidemiol Biomark Prevent* 2015; **24**: 1495–1506.

26. Fitzgibbons PL, Page DL, Weaver D, et al. Prognostic factors in breast cancer: College of American Pathologists consensus statement 1999. *Arch Pathol Lab Med* 2000; **124**: 966–978.

27. Berg WA, Gutierrez L, NessAiver MS, et al. Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer 1. *Radiology* 2004; **233**: 830–849.

28. Kattan WM, Giri D, Panageas KS, et al. A tool for predicting breast carcinoma mortality in women who do not receive adjuvant therapy. *Cancer* 2004; **101**: 2509–2515.