Estimating nonlinear effects in the presence of cure fraction using a semi-parametric regression model

Ramires, Thiago G.; HENS, Niel; Cordeiro, Gauss M. & Ortega, Edwin M. M. (2018)
Estimating nonlinear effects in the presence of cure fraction using a semi-parametric regression model. In: COMPUTATIONAL STATISTICS, 33(2), p. 709-730.

DOI: 10.1007/s00180-017-0781-8
Handle: http://hdl.handle.net/1942/26582

Estimating nonlinear effects in the presence of cure fraction using a semi-parametric regression model

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;
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
Edwin M.M. Ortega
Department of Exact Sciences, University of São Paulo, Brazil

Abstract

Nonlinear effects between explanatory and response variables are increasingly present in new surveys. In this paper, we propose a flexible four-parameter semi-parametric cure rate survival model called the sinh Cauchy cure rate distribution. The proposed model is based on the generalized additive models for location, scale and shape, for which any or all parameters of the distribution are parametric linear and/or nonparametric smooth functions of explanatory variables. The new model is used to fit the nonlinear behavior between explanatory variables and cure rate. The biases of the cure rate parameter estimates caused by not incorporating such non-linear effects in the model are investigated using Monte Carlo simulations. We discuss diagnostic measures and methods to select additive terms and their computational implementation. The flexibility of the proposed model is illustrated by predicting lifetime and cure rate proportion as well as identifying factors associated to women diagnosed with breast cancer.

Keywords: Cure rate models; GAMLSS; Long-term survivors; P-spline; Residual analysis.

1 Introduction

The objective of this study is to analyze censored data with the presence of long-duration individuals in which explanatory variables have nonlinear effects in relation to the failure times. Regression models with cure fraction are characterized by a significant fraction of individuals that do not experience the event of interest, even after a long follow-up period. In many cases, explanatory variables can present indefinite behavior. Nonlinear effects between explanatory and response variables are increasingly present in literature. A natural question that arises is how to deal with nonlinearity in the relationship between the outcome variable and a continuous predictor. The incorrect assumption of linearity can lead to a misspecified final model in which a relevant/irrelevant variable may not be included/excluded due to the fact that the hypothesis tests of the parameters related to such variables are based on the slope of the estimated line. Therefore, with the objective of obtaining a more flexible fit to the data, we use nonparametric functions to study the relationship between the response variable and the explanatory variables, allowing greater flexibility by not imposing a rigid dependence form in modeling the variables in question.

One possible solution would be use categorization, in which such predictors are entered into stepwise selection procedures as linear terms or as dummy variables obtained after grouping. To exemplify, we present in Figure 1(a) the empirical survival curves for the recurrence free survival times as functions of the explanatory variable *age*, categorized in three levels, $age < 35$, $35 \leq age \leq 55$ and $age > 55$. The description of this data set is presented in Section 5, in which a thorough study is conducted. Note that the the proportion of cured individuals increases and then decreases when age increases, indicating a nonlinear effect of *age* in the cure rate proportion. These effects of *age* in the cure rate proportion can be noted in Figure 1(b), where we display the fitted cure rate proportions for each category of *age* using nonparametric techniques.

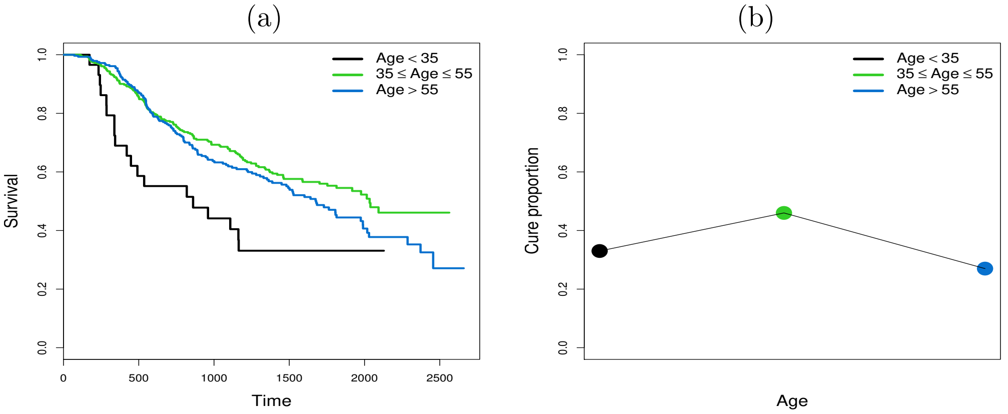

Figure 1: (a) The empirical survival curves as functions of the categorized explanatory variable *age* and (b) the estimated cure rate proportion obtained for each of its category.

The problem in the categorization method is that it introduces problems of defining cutpoints (Altman *et al.*, 1994), over-parametrization and loss of efficiency (Morgan and Elashoff, 1986; Lagakos, 1988). In any case, a cutpoint model is an unrealistic way to describe a smooth relationship between a predictor and an outcome variable, which will depend on the priori chosen by the researcher, and it may not be realistic. Nonparametric regression methods are alternative to parametric modelling of curved relationships. Some methods that have been emphasized in the statistical area are: regression splines, smoothing splines and kernel methods (Hastie and Tibshirani, 1990; Green and Silverman, 1993). Although these methods are relatively advanced, usually such techniques are only adopted on location-and-scale models, thus requiring the expansion of such techniques to other kinds of models like long-term survival.

In regression analysis, one or more explanatory variables can have significant effects on the location parameter, but also on other parameters such as scale and skewness parameters. The erroneous consideration of the regression structure can have adverse consequences for the efficiency of estimators, so it is important to consider the regression structure for all model parameters whenever possible. In this paper, we propose a general class of regression models with cure fraction, where the mean, dispersion, skewness (bi-modality) and cure fraction parameters vary across observations through regression structures. This type of model is called in literature as the *generalized additive model for location, scale and shape* (GAMLSS) (Rigby and Stasinopouls, 2005). We also consider, for each model parameter, smoothing techniques to capture nonlinear effects existent in the continuous explanatory variables.

For modeling a lifetime $T > 0$, the log-sinh Cauchy (LSC) distribution (Ramires *et al.*, 2016) was introduced to accommodate various shapes of skewness, kurtosis and bi-modality, being flexible for a wide range of data. We consider that the failure times follow the LSC distribution and propose a new model called the *log-sinh Cauchy cure rate* (LSCcr) model. The paper is organized as follows. In Section 2, we define the LSCcr model by means of the density and survival functions. Further, we propose the *log-sinh Cauchy cure rate generalized additive model for location, scale and shape* (LSCcr GAMLSS) and discuss about smooth functions. Inferential issues, model selection strategies, goodness-of-fit, selection of the additive terms and residual analysis are investigated in Section 3. In Section 4, we discuss methods for generating random values and provide Monte Carlo simulations on the finite sample behavior of the maximum likelihood estimates (MLEs). An application to breast cancer data presented in Section 5 illustrates the flexibility of the proposed semi-parametric regression model. Computational implementation and instructions for fitting the new model are given in the Appendix. Finally, we offer some conclusions in Section 6.

2 The LSC semi-parametric regression model with long-term survivors

Models to accommodate a cured fraction have been widely developed. The literature on the subject is by now rich and growing rapidly. The books by Maller and Zhou (1996) and Ibrahim*et al.* (2001) as well as the review papers by Chen *et al.* (1999), Tsodikov *et al.* (2003) and the article by Cooner *et al.* (2007) could be mentioned as key references. Currently, several works have been developed considering cure rate models, e.g., Balakrishnan and Pal (2012) pioneered an EM algorithm-based likelihood estimation for some cure rate models, Cancho *et al.* (2015) studied a unified multivariate survival model with a surviving fraction, Hashimoto *et al.* (2015) defined a new long-term survival model with interval-censored data, Ortega *et al.* (2015) defined a power series beta Weibull regression model for predicting breast carcinoma and Cordeiro *et al.* (2016) proposed the negative binomial Birnbaum-Saunders model with long-term survivors.

Perhaps the most popular type of cure rate models are the mixture models (MMs) defined by Boag (1949), Berkson and Gage (1952) and further studied by Farewell (1982). This approach allows simultaneously estimating whether the event of interest will occur, which is called incidence, and when it will occur, given that it can occur, which is called latency. Let N_i (for $i = 1, \ldots, n$) be the indicator denoting that the ith individual is susceptible to failure (uncured) ($N_i = 1$) or non-susceptible (cured) ($N_i = 0$), i.e., the population is classified in two sub-populations so that an individual either is cured with probability $0 < \tau < 1$, or has a proper survival function $S(t)$ with probability $(1 - \tau)$. The MM can be expressed as

$$S_{pop}(t_i) = \tau + \big(1 - \tau\big) S(t_i|N_i = 1), \tag{1}$$

where $S_{pop}(t_i)$ is the unconditional survival function of t_i for the entire population, $S(t_i|N_i = 1)$ is the survival function for susceptible individuals and $\tau = P(N_i = 0)$ is the probability of cure of an individual. The probability density function (pdf) corresponding to (1) is given by

$$f_{pop}(t_i) = -\frac{d\,S_{pop}(t_i)}{dt} = (1 - \tau)\, f(t_i|N_i = 1), \tag{2}$$

where $f(t_i|N_i = 1)$ is the baseline pdf for the susceptible individuals. Equations (1) and (2) are improper functions, since $S_{pop}(t)$ is not a proper survival function. We omit the dependence on the indicator N_i and write simply $S(t_i|N_i = 1) = S(t)$, $f(t_i|N_i = 1) = f(t)$, etc.

Recently, for modeling a lifetime $T > 0$, Ramires *et al.* (2016) introduced the LSC distribution, which accommodates various shapes of the skewness, kurtosis and bi-modality. Its density function is given by

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

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

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

2.1 The LSCcr distribution

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. We define the LSCcr 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).

The survival function for the LSCcr model is defined by assuming that the survival function for susceptible individuals in (1) is given by (4), which gives

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

where $w = \frac{\log(t)-\mu}{\sigma}$. We again omit the dependence on the parameters as, for example, $S_{pop}(t) = S_{pop}(t;\mu,\sigma,\nu,\tau)$. The pdf corresponding to (5) is given by

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

The hazard rate function (hrf) of the LSCcr model is given by $h_{pop}(t) = f_{pop}(t)/S_{pop}(t)$. A random variable having density (6) is denoted by $T \sim LSCcr(\mu,\sigma,\nu,\tau)$. Clearly, the functions $f_{pop}(t)$ and $h_{pop}(t)$ are improper functions, since $S_{pop}(t)$ is not a proper survival function. Plots of the LSCcr survival and hazard functions for selected parameter values are displayed in Figures 2 and 3, respectively.

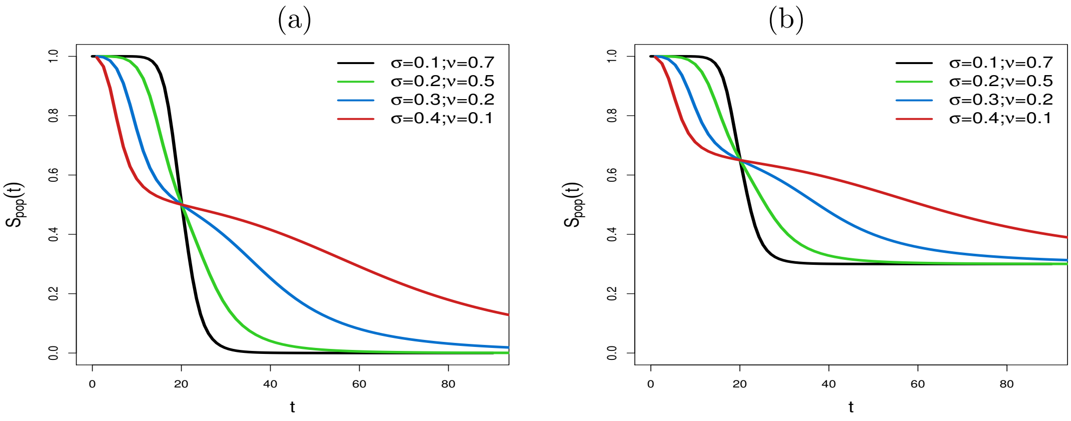

Figure 2: The LSCcr survival function when $\mu = 3$ and: (a) For $\tau = 0$ and different values of σ and ν; (b) For $\tau = 0.3$ and different values of σ and ν.

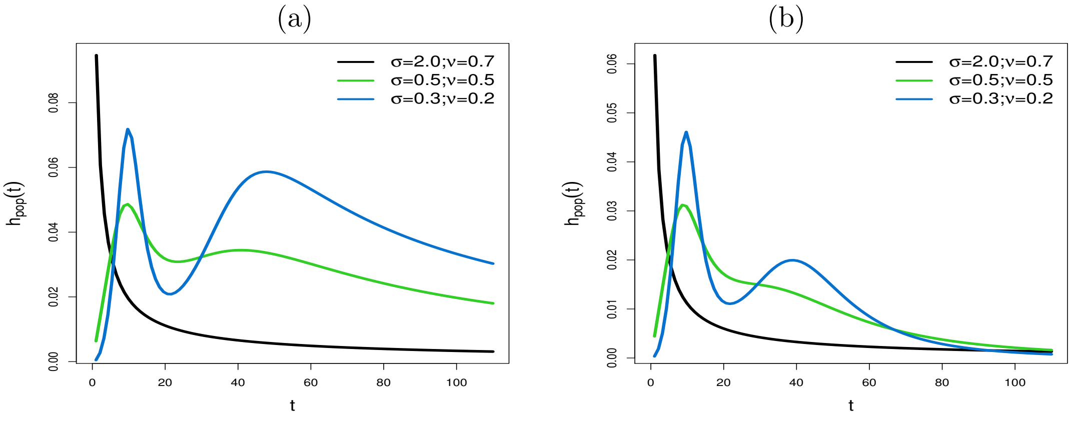

Figure 3: The LSCcr hrf when $\mu = 3$ and: (a) For $\tau = 0$ and different values of σ and ν; (b) For $\tau = 0.3$ and different values of σ and ν.

Figures 2(a)-(b) clearly reveal the symmetric and bi-modality effects due to the parameters σ and ν, respectively, and different effects of the cured probability τ. Further, Figures 3(a)-(b) indicate that the hrf of T can have decreasing, unimodal and bimodal shapes. We can note in Figure 3(b) that the values of the hrf are smaller in the presence of the proportion of cured individuals but still assuming the same characteristics.

2.2 The LSCcr semi-parametric regression model

In many practical applications, the response variables are affected by explanatory variables. In the presence of explanatory variables with nonlinear effects, semi-parametric models are widely used. If these models provide good fits, they tend to give more precise estimates of the quantities of interest. Recently, several regression models have been proposed in literature by considering the class of location models. For example, Ortega *et al.* (2014) introduced a log-linear regression model for the odd Weibull distribution, da Cruz *et al.* (2016) proposed the log-odd log-logistic Weibull regression model with censored data, Lanjoni *et al.* (2016) studied the extended Burr XII regression models and Hashimoto *et al.* (2016) defined a new flexible regression model generated by gamma random variables with censored data. A disadvantage of the class of location models is that the variance and skewness and other parameters are not modelled explicitly in terms of explanatory variables but only implicitly through their dependence on the location parameter. As an alternative, the GAMLSS (Rigby and Stasinopouls, 2005) allows all parameters of the conditional distribution of T be modelled as functions of the explanatory variables.

On the other hand, in most studies considering regression models, the structure of continuous covariates is added in the models such that it is linear in the parameters regarding the proportion of cured individuals, although this relationship is not always true. The inappropriateness of the structures of the regression models makes it impossible to capture the variability of such covariates in the model, degrading the estimates of all other parameters to be estimated, and in the worst cases, leading to the wrong conclusion that these variables do not have significant effects on cure rates. To capture the nonlinear effects of these covariates, it is necessary to adopt nonlinear functions.

Let $T \sim LSCcr(y;\theta)$, where $\theta = (\mu,\sigma,\nu,\tau)^T$ denotes the vector of parameters of the pdf (6). Consider independent observations t_i conditional on the parameter vector $\boldsymbol{\theta}_i$ (for $i = 1,\dots,n$) having pdf $f(t_i;\boldsymbol{\theta}_i)$, 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. The GAMLSS allows the user to model all parameters in $\boldsymbol{\theta}$ as linear, nonlinear parametric, nonparametric (smooth) function of the explanatory variables and/or random effects terms. We can define semi-parametric structures for the elements of the vector $\boldsymbol{\theta}$ using appropriate link functions as

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\sigma} \\ \boldsymbol{\nu} \\ \boldsymbol{\tau} \end{bmatrix} = \begin{bmatrix} g_1\Big(\mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1})\Big) \\ g_2\Big(\mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} h_{j2}(\mathbf{x}_{j2})\Big) \\ g_3\Big(\mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} h_{j3}(\mathbf{x}_{j3})\Big) \\ g_4\Big(\mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} h_{j4}(\mathbf{x}_{j4})\Big) \end{bmatrix}, \tag{7}$$

where $g_k(\cdot)$ for $k = 1,2,3,4$ denote the injective and twice continuously differentiable monotonic link functions, $\boldsymbol{\beta}_k = (\beta_{0k},\beta_{1k},\dots,\beta_{m_k k})^T$ 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 $n \times (m_k+1)$. Here, $h_{jk}(\mathbf{x}_{jk})$ are smooth functions of the explanatory variables $\mathbf{x}_{jk}$ for $j = 1,\dots,J_k$. The explanatory variables can be similar or different for each of the distribution parameters, which can be considered as linear functions, smooth functions or both. In the following sections, we shall consider the identity link function for $g_1(\cdot)$, the logarithmic link function for $g_k(\cdot)$ $(k = 2,3)$ and the logit link function for $g_4(\cdot)$.

In this paper, we only use the P-splines as smooth functions $h_{jk}(\cdot)$. The P-splines are piecewise polynomials defined by B-spline basis functions in the explanatory variables, where the coefficients of the basis functions are penalized to guarantee sufficient smoothness. Rigby and Stasinopouls (2005) have shown that each smoothing function $h_{jk}(\cdot)$ can be expressed as a random effects model, i.e., $h_{jk}(\cdot) = \mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}$, where $\mathbf{Z}_{jk}$ is an $n \times q_{jk}$ matrix representing the B-spline basis design matrix and $\boldsymbol{\gamma}_{jk}$ is a q_{jk}-dimensional vector of the B-spline parameters (random-effects). Some details of the number of knots and the degrees of freedom can be found in Eilers and Marx (1996).

3 Model selection

In this section, we present the numerical maximization methods to fit the LSCcr GAMLSS and some procedures to select the best model and additive terms as well as some diagnostic techniques.

3.1 Inference

The numerical maximization of the log-likelihood can be performed in the `GAMLSS` and `gamlss.cens` packages of the `R` software using the computational codes implemented by the first author. The maximization algorithms used are the RS and CG procedures described by Rigby and Stasinopouls (2005) and Stasinopoulos and Rigby (2007) and available in the documentation of the GAMLSS package.

Consider a sample of n-independent observations $t_1, \cdots, t_n$, noninformative censoring and that the observed lifetimes and censoring times are independent. Let F and C be the sets of individuals for which t_i is the lifetime or censoring, respectively. For the semi-parametric model (7), the fixed and random effects $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively, are estimated by maximizing the penalized log-likelihood function

$$l_p = l(\boldsymbol{\theta}) - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \lambda_{jk}\boldsymbol{\gamma}_{jk}^T \mathbf{P}_{jk}\,\boldsymbol{\gamma}_{jk}, \qquad (8)$$

where $\mathbf{P}_{jk}$ is a symmetric matrix that may depend on a vector of smoothing parameters, see for example, Rigby and Stasinopouls (2005). For each smoothing term selected, and any of the parameters of the LSCcr distribution, there is one smoothing parameter λ associated with it. The smoothing parameters can be fixed or estimated from the data. We adopt the Penalised Quasi Likelihood (PQL) method, described by Lee *et al.* (2006), to estimate the smoothing parameters and the degrees of freedom of the P-spline smooth functions. This method is implemented in the `R` software in the `pb(.)` function (Rigby and Stasinopouls, 2014). One important thing to remember when fitting a smooth nonparametric term is the fact that the resulting coefficients of the smoothing terms and their standard errors should not be interpreted.

The non-penalized log-likelihood function $l(\boldsymbol{\theta}) = \sum_{i \in F} \log f_{pop}(t_i; \boldsymbol{\theta}_i) + \sum_{i \in C} \log S_{pop}(t_i; \boldsymbol{\theta}_i)$ is given by

$$
\begin{aligned}
l(\boldsymbol{\theta}) \;=\; & \sum_{i\in F}\Big\{\log(1-\tau_i)+\log(\nu_i)-\log(\sigma_i\pi)-\log(t_i)+\log\cosh(w_i)-\log\left[1+\nu_i^2\,\sinh^2(w_i)\right]\Big\} \\
& +\sum_{i\in C}\log\left(1+(p_i-1)\left\{\frac{1}{2}+\frac{1}{\pi}\arctan\left[\nu_i\,\sinh\left(w_i\right)\right]\right\}\right), \qquad (9)
\end{aligned}
$$

where $w_i=[\log(t_i)-\mu_i]/\sigma_i$. The parameter vector $\boldsymbol{\theta}=(\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_4^T)^T$ is used to define the regression structures in (7) by specifying appropriate link functions for $g_k(\cdot)$, e.g., using the logit link function for $g_4(\boldsymbol{\tau})$, the parameter τ is related to the covariates by $\tau_i=\exp(\mathbf{X}_4[i,]\boldsymbol{\beta}_4)/[1+\exp(\mathbf{X}_4[i,]\boldsymbol{\beta}_4)]$, where $\mathbf{X}_k[i,]$ denotes the i-th row of the design model matrix $\mathbf{X}_k$. The fit of the LSCcr model gives the vector of estimated cured proportion

$$
\hat{\boldsymbol{\tau}}=\frac{\exp[\mathbf{X}_4\hat{\boldsymbol{\beta}}_4+\sum_{j=1}^{J_4}\hat{h}_{j4}(\mathbf{x}_{j4})]}{1+\exp[\mathbf{X}_4\hat{\boldsymbol{\beta}}_4+\sum_{j=1}^{J_4}\hat{h}_{j4}(\mathbf{x}_{j4})]}, \qquad (10)
$$

where $\hat{h}_{j4}(\mathbf{x}_{j4})=\mathbf{Z}_{j4}\hat{\boldsymbol{\gamma}}_{j4}$.

Let df_μ, df_σ, df_ν and df_τ be the effective degrees of freedom used for modelling μ, σ, ν and τ, respectively. The df combines the effective degrees of freedom used in the smooth functions $h_{jk}(\cdot)$ and parametric functions defined by $df=df_\mu+df_\sigma+df_\nu+df_\tau$. For example, let the location parameter be modelled by the explanatory variable X_1 using a nonparametric smoothing function with five additional degrees of freedom. Then, the effective degrees of freedom related to the location parameter is given by $df_\mu=5+2$, where the additional two degrees of freedom account for the linear term. The effective degrees of freedom related to the smoothing function are defined by the trace of the corresponding smoothing matrix in the fitting algorithm, which is in turn directly related to the corresponding smoothing parameter (Eilers and Marx, 1996). The df can be evaluated using the `edfAll(.)` function in the R software.

3.2 Goodness-of-fit

The selection of the appropriate distribution is performed in two stages, the fitting stage and the diagnostic stage. In the first stage, we use the global deviance (GD), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The GD is given by $GD=-2\,l_p(\hat{\boldsymbol{\theta}})$, where $l_p(\hat{\boldsymbol{\theta}})$ is the total log-likelihood function and the AIC and BIC criterion are obtained by $AIC=GD+2\,df$ and $AIC=GD+\log(n)\,df$, where df is the total effective degrees of freedom of the fitted model. The model with the smallest values of these criteria is then selected.

In the diagnostic stage, the model assumptions and the presence of outlying observations are checked. We can use the diagnostic tools in the GAMLSS package. The first technique consists in the normalized randomized quantile residuals (Dunn and Smyth, 1996), 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 distribution, $\hat{u}_i=1-S_{pop}(t_i|\hat{\boldsymbol{\theta}}_i)$ and $S_{pop}(t_i|\hat{\boldsymbol{\theta}}_i)$ is the survival function (5). For censored observations, considering a right censored continuous response, $\hat{u}$ is defined as a random value from a uniform distribution on the interval $[1-S_{pop}(t_i|\hat{\boldsymbol{\theta}}_i)\,,1]$.

The second technique involves the use of Worm Plots (WP). These plots of the residuals were pioneered by Buuren and Fredriks (2001) in order to identify regions (intervals) of an explanatory variable within which the model does not adequately fit the data. This is a diagnostic tool for checking the residuals for different ranges of one or two explanatory variables. Buuren and Fredriks (2001) proposed fitting 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.

3.3 Additive terms selection

For the LSCcr GAMLSS, the selection of the terms for all the parameters is performed using the stepwise generalized Akaike information criterion (GAIC) procedure. There are many different strategies that could be applied for the selection of the terms used to model the four parameters μ, σ, ν and τ. Here, we consider a modification of the strategy described by Voudouris *et al.* (2012). Let χ be the selection of all terms available for consideration, where χ could contain both linear and smoothing terms. Then, for all terms in χ and for fixed distribution, the strategy is given as follows (we suggest to use of the AIC criterion for the next steps):

- Use the forward produce, in which each covariate that is not already in model is tested for inclusion, to select the additive terms for the parameter τ considering μ, σ and ν fixed (without covariates);
- Considering the model selected for τ, use the forward produce to select the additive terms for μ after σ and then for ν, while fixing the model obtained in the previous step.

By the end of the steps described above, the final model may contain different subsets from χ for $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$.

4 Simulation study

Consider the random variable T having pdf (3). By inverting $F(t) = 1 - S(t) = u$ in (4), we obtain the qf of the LSC distribution as

$$t = Q(u) = \exp\left(\mu + \sigma \operatorname{arcsinh}\left\{\frac{1}{\nu}\tan\left[\pi\left(u - 0.5\right)\right]\right\}\right). \tag{11}$$

Equation (11) can be used for simulating $T \sim \mathrm{LSC}(\mu, \sigma, \nu)$ by fixing the parameters μ, σ and ν and setting u as a uniform random variable in the interval $(0, 1)$. The cured proportion can be generated using the qf of another distribution with real support, fixing τ and setting the sample size for the cured individuals as $n_c = \tau \times n$, where n denotes the total sample size. We can also simulate the regression models by setting the parameters using the semi-parametric (7) structure.

Here, we conduct a Monte Carlo simulation study, considering two scenarios, to assess the finite sample behavior of the MLEs of the model parameters. We consider model (7), where the cure rate parameter τ has a nonlinear relationship with the explanatory variables X_1 and X_2, related to scenarios 1 and 2, respectively. The total sample sizes is taken as $n = 100, 200$ and 300 and parameters values are fixed at $\mu = 2.5$, $\sigma = 0.5$ and $\nu = 0.5$. The values of the parameter τ are defined such that X_1 and X_2 has a quadratic and cubic effect in τ, respectively. For each level of X_1 and X_2, a sample of size $n/10$ was generated. The fixed values of τ, for each value of X_1 and X_2, are given in Table 1. The total of censoring percentages are 42% and 35,5% for scenarios 1 and 2, respectively.

Table 1: Fixed values of the τ parameter for each level of the X_1 and X_2 explanatory variables.

Quadratic	τ	0.20	0.35	0.40	0.55	0.60	0.60	0.55	0.40	0.35	0.20
effect	X_1	1	2	3	4	5	6	7	8	9	10
Cubic	τ	0.10	0.30	0.40	0.45	0.30	0.25	0.25	0.35	0.40	0.75
effect	X_2	1	2	3	4	5	6	7	8	9	10

The failure times T, denoted by $t_1, \cdots, t_n$, are generated from the LSC distribution using the qf (11) and the censoring times C are randomly generated from the uniform distribution $C \sim U[\max(T), \max(T) + 2\, sd(T)]$, where $sd(T)$ denotes the standard deviation of the failure time sample. The lifetimes considered in each fit are evaluated as $\min(t_i; c_i)$, where all results are obtained from 1,000 Monte Carlo replications. For each replication, we evaluate the MLEs of the parameters and then, after all replications, we compute the average estimates (AEs), biases and means squared errors (MSEs).

Next, we present and compare the results by fitting the parametric and semi-parametric LSCcr models, for both scenarios, namely

- **Parametric** LSCcr$(\mu, \sigma, \nu, \text{logit}[\beta_{04} + \beta_{14}\, X_k])$, for $k = 1, 2$,
- **Semi-parametric** LSCcr$(\mu, \sigma, \nu, \text{logit}[pb(X_k, df)])$, for $k = 1, 2$,

where $pb(X_k, df)$ denotes a smooth P-spline function with corresponding degrees of freedom df to model the dependence on X_1 or X_2. The purpose of this study is to compare the loss of efficiency caused by a misspecified model. The AEs, biases and MSEs are evaluated and the results for both scenarios are reported in Table 2. We also present the averages of the AIC, BIC and GD statistics obtained in the 1,000 simulations. As the coefficients of the smoothing terms $pb(X_k, df)$ are meaningless, we only present averages of the estimated degrees of freedom in this table.

The figures in Table 2 reveal that the MSEs of the MLEs of the parameters for the parametric and semi-parametric models are very close. Note that the average of the effective degree of freedom df for the semi-parametric models is not close to two, thus indicating that we have a significative nonlinear effect of X_1 and X_2 in the cure rate parameter. By taking into account the parameter estimates relative to the cure rate parameter for the parametric models, we note that β_{14} is approximately zero for scenario 1, erroneously indicating that the explanatory variables X_1 and X_2 has no effect on the cure rate proportions. The main conclusion of this simulation study is that, when the regression model is specified incorrectly, i.e., not allowing that nonlinear effects can be estimated, erroneous conclusions can be drawn about the explanatory variables. We can also conclude that the BIC statistics is not

Table 2: The AEs, biases, MSEs and the averages of the goodness-of-fit statistics for the parametric and semi-parametric LSCcr regression models based on 1,000 simulations for scenarios 1 and 2.

		Parametric				Semi-parametric			
k	n	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
1	100	μ	2.817	0.317	0.151	μ	2.765	0.265	0.105
		σ	0.589	0.089	0.025	σ	0.577	0.077	0.022
		ν	0.501	0.001	0.044	ν	0.504	0.004	0.044
		β_{04}	-0.989	-	-	df	2.703	-	-
		β_{14}	0.017	-	-				
		$AIC = 572.7$; $BIC = 585.7$; $GD = 562.7$				$AIC = 564.8$; $BIC = 579.6$; $GD = 553.4$			
1	200	μ	2.657	0.157	0.038	μ	2.632	0.132	0.028
		σ	0.578	0.078	0.013	σ	0.571	0.071	0.012
		ν	0.542	0.042	0.023	ν	0.544	0.044	0.023
		β_{04}	-0.563	-	-	df	3.156	-	-
		β_{14}	0.000	-	-				
		$AIC = 1174.1$; $BIC = 1190.5$; $GD = 1164.1$				$AIC = 1163.1$; $BIC = 1182.9$; $GD = 1151.0$			
1	300	μ	2.604	0.104	0.017	μ	2.594	0.094	0.015
		σ	0.570	0.070	0.010	σ	0.565	0.065	0.009
		ν	0.563	0.063	0.020	ν	0.561	0.061	0.019
		β_{04}	-0.496	-	-	df	3.215	-	-
		β_{14}	0.000	-	-				
		$AIC = 1782.0$; $BIC = 1800.5$; $GD = 1772.0$				$AIC = 1763.6$; $BIC = 1786.6$; $GD = 1751.2$			
2	100	μ	2.824	0.324	0.169	μ	2.744	0.244	0.092
		σ	0.588	0.088	0.023	σ	0.566	0.066	0.018
		ν	0.497	-0.002	0.044	ν	0.497	-0.002	0.043
		β_{04}	-2.462	-	-	df	3.531	-	-
		β_{14}	0.173	-	-				
		$AIC = 589.2$; $BIC = 602.2$; $GD = 579.2$				$AIC = 583.5$; $BIC = 600.5$; $GD = 570.4$			
2	200	μ	2.642	0.142	0.031	μ	2.631	0.131	0.027
		σ	0.574	0.074	0.012	σ	0.569	0.069	0.011
		ν	0.548	0.048	0.023	ν	0.546	0.046	0.022
		β_{04}	-1.843	-	-	df	2.947	-	-
		β_{14}	0.169	-	-				
		$AIC = 1257.7$; $BIC = 1274.2$; $GD = 1247.7$				$AIC = 1252.7$; $BIC = 1272.3$; $GD = 1240.8$			
2	300	μ	2.590	0.090	0.014	μ	2.580	0.080	0.012
		σ	0.564	0.064	0.008	σ	0.559	0.059	0.007
		ν	0.559	0.059	0.018	ν	0.557	0.057	0.017
		β_{04}	-1.631	-	-	df	3.687	-	-
		β_{14}	0.154	-	-				
		$AIC = 1886.8$; $BIC = 1905.3$; $GD = 1876.8$				$AIC = 1874.6$; $BIC = 1899.4$; $GD = 1861.2$			

appropriated to compare parametric with semiparametric models due to high penalty in the number of parameters.

Figure 4 displays the generated and fitted effects for the parametric and semi-parametric models, considering $n = 200$, for both scenarios. We also present in this figure the box-plots of the GD, AIC and BIC statistics obtained in 1,000 simulations. We can note that the estimates of the cure rate parameter $\hat{\tau}$ are more suitable for the semi-parametric models. Further, we can conclude that the semi-parametric models present the lowest values of the GD, AIC and BIC statistics, thus indicating to be the most appropriate models.

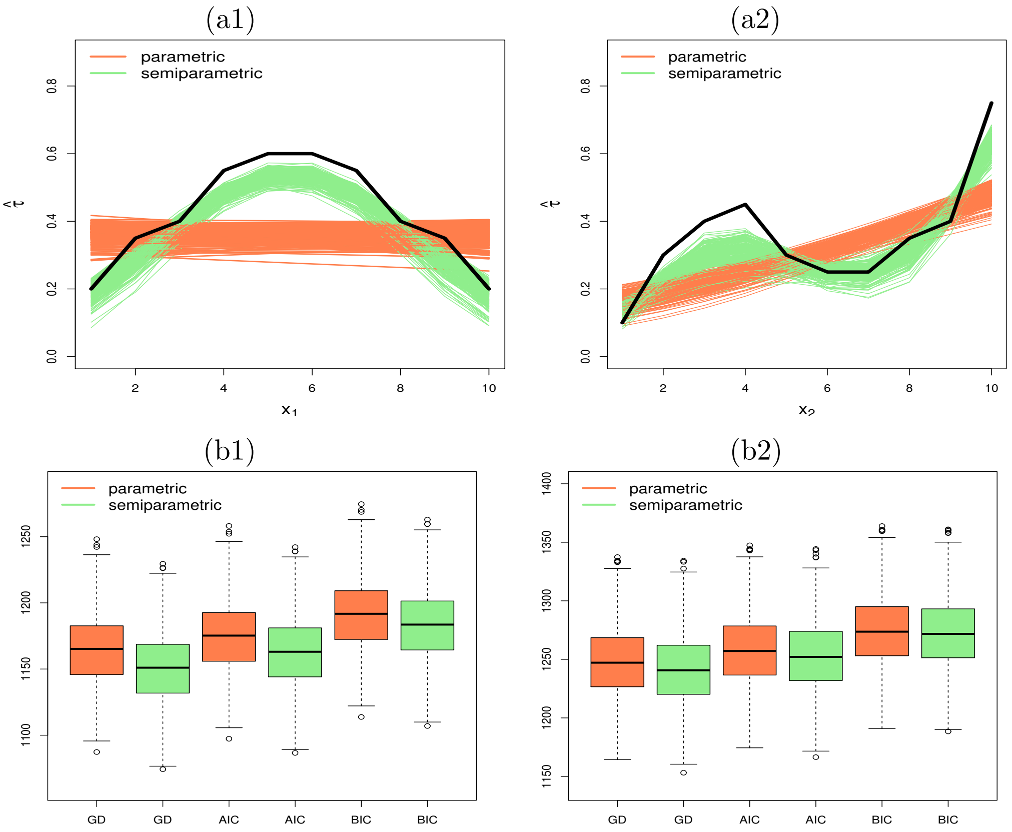

Figure 4: For the fitted LSCcr parametric and semi-parametric models, for $n = 200$: the true and fitted effect of (a1) X_1 and (a2) X_2 in the parameter τ and the goodness-of-fit statistics for scenarios (b1) 1 and (b2) 2.

5 Predicting the cure rate of breast cancer

A prognosis is the physician best estimate of how cancer will affect a person. A predictive factor influences how a cancer will respond to a certain treatment. Prognostic and predictive factors are often discussed together and they both play a significant part in deciding on a treatment plan and a prognosis. The following are prognostic and predictive factors for breast cancer.

The initial prognostic model considers that the explanatory variables tumor size, histology grade, and lymph node status as basic factors to be taken into account (Fitzgibbons *et al.*, 2000). A woman's age at the time of her breast cancer diagnosis can affect the prognosis. Younger women (under 35 years of age) usually have a greater risk of recurrence. The size of a breast tumor is the second most important prognostic factor for breast cancer, in which the size of the tumor increases the risk of recurrence. The grade of the breast cancer also affects prognosis, low-grade tumors often grow slower and are less likely to spread than high-grade tumors (Gospodarowicz *et al.*, 2006; Ko, 2009; Lønning, 2007).

In this section, we predict disease-free survival time (death, second malignancy or cancer recurrence considered as event) by means of a data set corresponding to women diagnosed with breast cancer in Germany (Schumacher *et al.*, 1994). The data comprises 686 node positive women who had complete data for these predictors. These women experienced 299 (43.6%) events during a median follow-up time of 53.9 months, leaving all other patients with a right censored failure time.

The explanatory variables measures in the study are described below:

- t_i: recurrence free survival time (in days);
- δ_i: failure indicator (0: censored, 1: observed);
- *age*: age (in years);
- *htreat*: hormonal treatment with tamoxifen (0: no, 1: yes);
- *menostat*: menopausal status (1: premenopausal, 2: postmenopausal);
- *tumsize*: tumor size (in mm);
- *tumgrad*: tumor grade, a ordered factor at levels $(1 < 2 < 3)$;
- *posnodal*: number of positive lymph nodes;
- *prm*: progesterone receptor (in fmol);
- *esm*: estrogen receptor (in fmol).

We start the analysis describing the explanatory variables. Figure 5 displays the empirical survival functions and the corresponding p-values of the log-rank tests for the categorical variables. We may observe in these plots that only menopausal status does not present a significative difference between the survival curves. We also present the frequency histogram of three explanatory variables, progesterone receptor, tumor size and age, in Figure 6. These plots reveal that the highest concentration of progesterone receptor is in the range [0,600], the average tumor size is 29.3 and the average age is 53.

Next, using the steps described in Section 3.3 to select the additive terms for the different parameters, we provide results for the LSCcr GAMLSS parameters. We also compare the results by fitting the Weibull cure rate (Weibullcr) model with scale $\mu > 0$, shape $\sigma > 0$ and cure rate $\nu \in [0, 1]$ parameters. The model parameters are defined by

LSCcr model

$$\begin{aligned}
&\mu_i = \beta_{01} + \beta_{11} age + \beta_{21} prm + \beta_{31} htreat_1,\\
&\sigma_i = \exp(\beta_{02} + \beta_{12} tumgrad_2 + \beta_{22} tumgrad_3),\\
&\nu_i = \exp(\beta_{03}),\\
&\tau_i = \text{logistic}[\beta_{04} + \beta_{14} prm + \beta_{24} tumsize + \beta_{34} htreat_1 + \beta_{44} tumgrad_2 + \beta_{54} tumgrad_3 + pb(age)];
\end{aligned} \tag{12}$$

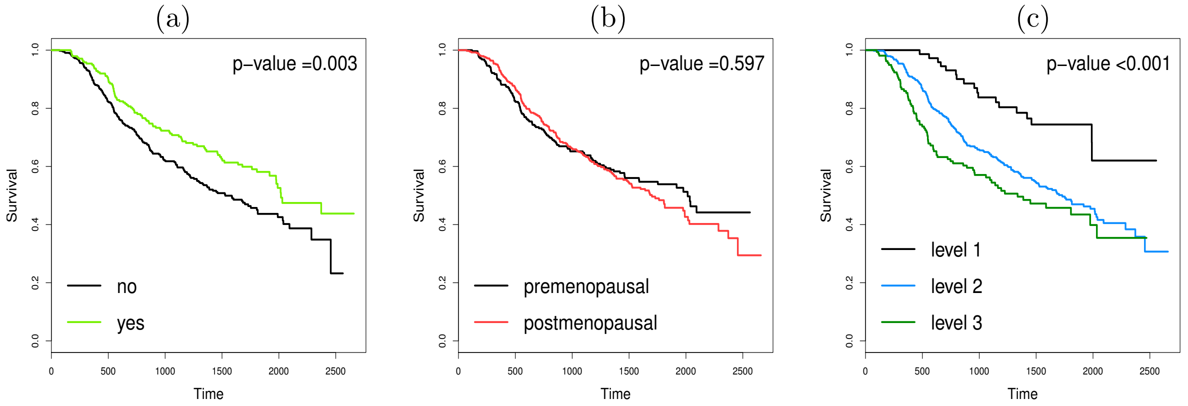

Figure 5: Empirical survival functions and log-rank tests for (a) *htreat* (b) *menostat* and (c) *tumgrad*.

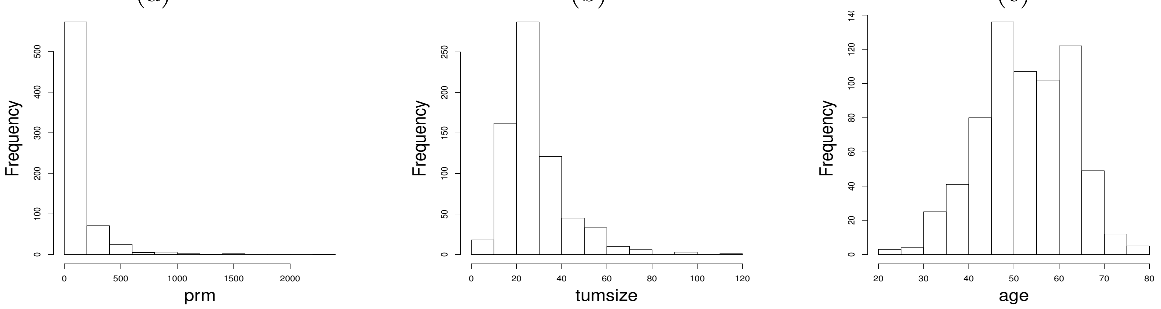

Figure 6: Frequency histogram of explanatory variables (a) progesterone receptor, (b) tumor size and (c) age.

Weibullcr model

$$\mu_i = \exp[\beta_{01} + \beta_{11} tumgrad_2 + \beta_{21} tumgrad_3 + pb(esm) + pb(age)],$$
$$\sigma_i = \exp(\beta_{02} + \beta_{12} tumgrad_2 + \beta_{22} tumgrad_3 + \beta_{32} tumsize),$$
$$\nu_i = \text{logistic}[\beta_{04} + \beta_{14} prm + \beta_{24} tumsize + \beta_{34} htreat_1 + \beta_{44} tumgrad_2 + \beta_{54} tumgrad_3 + pb(age)],$$

where $\text{logistic}(x) = \exp(x)/[1 + \exp(x)]$ and $htreat_1$, $tumgrad_2$ and $tumgrad_3$ are the indicator variables of $htreat = 1$, $tumgrad = 2$ and $tumgrad = 3$, respectively. Table 3 lists the values of the GD, AIC and BIC statistics for the fitted models. We can conclude from these figures that the LSCcr model provides a better fit than the Weibullcr model.

Table 3: The GD, AIC and BIC statistics and corresponding degrees of freedom for the fitted LSCcr and Weibullcr models.

Model	df	GD	AIC	BIC
LSCcr	18.18	5116.00	5152.37	5234.75
Weibullcr	27.81	5125.57	5181.20	5307.23

Table 4 provides the MLEs, SEs and p-values obtained from the fitted LSCcr GAMLSS. The coefficients of the smoothing terms have been omitted (to avoid misinterpretations). We may note in this table that all parameters are significant at 5%, indicating the efficiency of the selection method. We conclude that the explanatory variables *age*, *prm* and *htreat* are significant to fit the location parameter, only *tumgrad* is significant to explain the variability on t_i and *prm*, *tumsize*, *htreat*, *tumgrad* and *age* are significant to fit the cure rate parameter, where *age* has a nonlinear effect on it.

Table 4: MLEs of the parameters, approximate SEs and p-values from the fitted LSCcr GAMLSS.

Parameter	Estimate	SE	p-value	Parameter	Estimate	SE	p-value
β_{01}	6.223	0.126	<0.001	β_{04}	3.224	0.688	<0.001
β_{11}	0.0101	0.002	<0.001	β_{14}	0.002	0.001	<0.001
β_{21}	0.0007	0.001	<0.001	β_{24}	-0.046	0.010	<0.001
β_{31}	0.194	0.053	<0.001	β_{34}	0.519	0.208	0.012
β_{02}	-1.408	0.039	<0.001	β_{44}	-1.319	0.212	<0.001
β_{12}	0.306	0.046	<0.001	β_{54}	-1.678	0.351	<0.001
β_{22}	0.614	0.061	<0.001	$pb(age)$	$df = 5.183$		
β_{03}	-0.961	0.053	<0.001				

The partial effects of the explanatory variables in the location parameter μ are presented in Figure 7. From the model for μ, we may note that the recurrence free survival time t_i increases according to age (Panel (a)) and the progesterone receptor (Panel (b)) increases and it is greater for patients treated with hormonal treatment with tamoxifen (Panel (c)). Regarding the scale parameter σ, as we can see in Figure 8(a), the variability of t_i increases when the gradient tumor grade increases. For the cure rate parameter τ, we may conclude from Figure 8(b)-(f) that the probability of cure increases when the progesterone receptor increases, decreases as tumor size increases, is greater for patients who received hormonal treatment with tamoxifen, is higher for patients diagnosed with tumor grade 1 and is higher for patients age around 45 years.

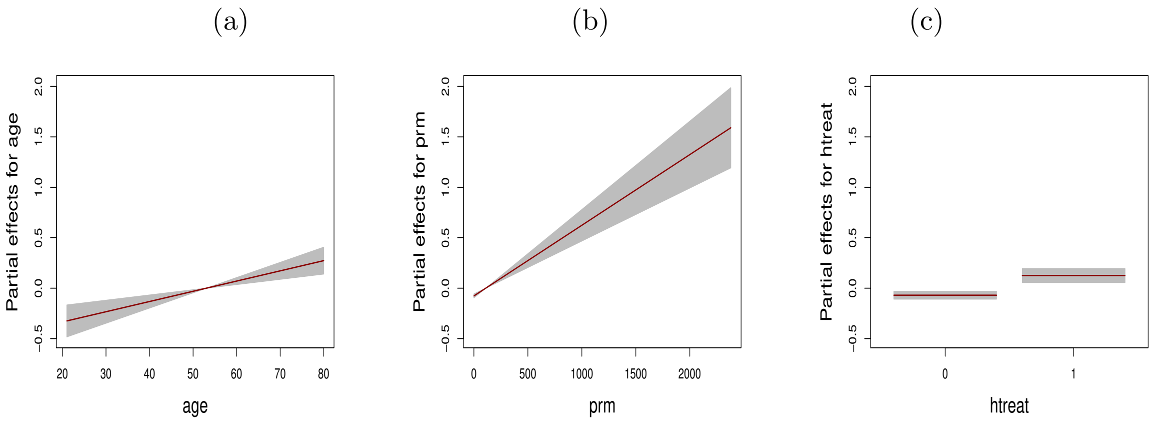

Figure 7: Fitted terms for the location μ parameter: (a) age, (b) progesterone receptor and (c) hormonal treatment.

Based on equation (10), the estimated cured proportions can be determined using the results obtained in (4) as $\hat{\tau}_i = \text{logistic}[3.224 + 0.002 prm_i - 0.046 tumsize_i + 0.519 htreat_{1i} - 1.319 tumgrad_{2i} -$

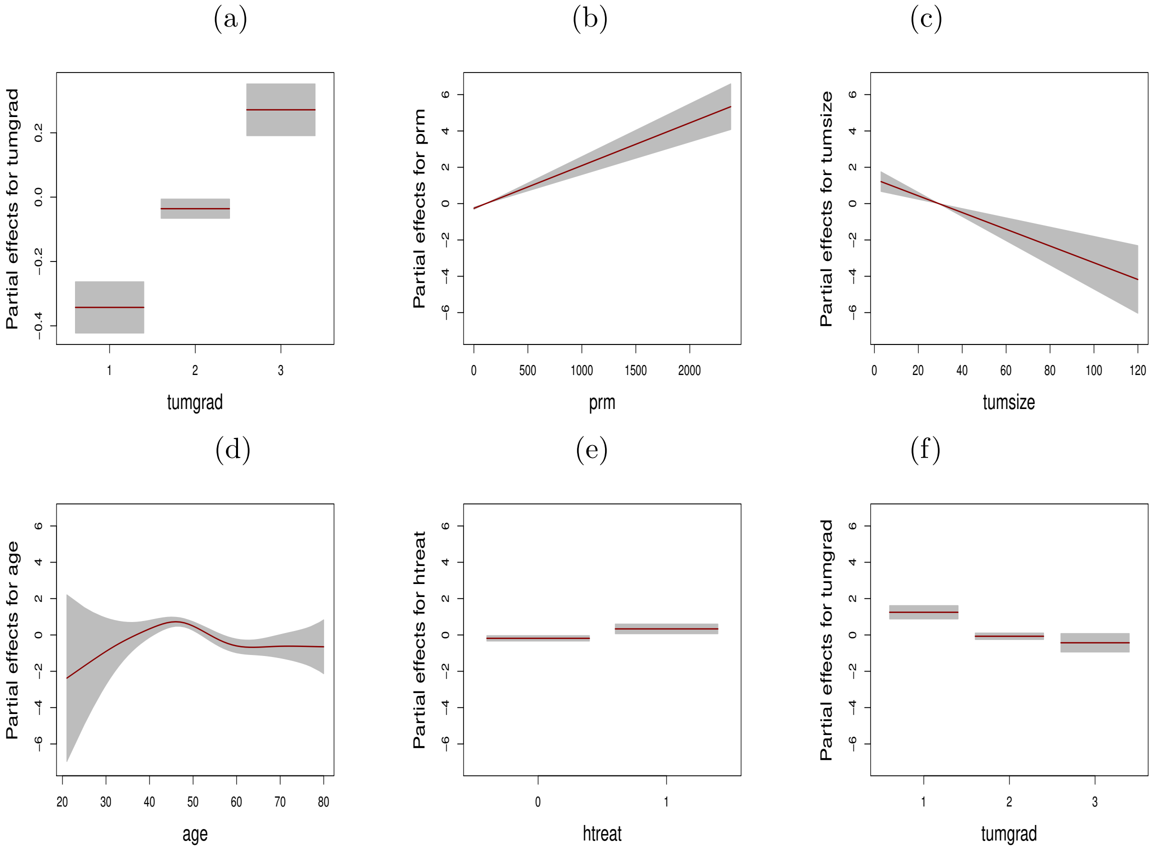

Figure 8: Fitted terms for (a) tumor grade covariate in the scale parameter σ, and for cure rate parameter τ, the fitted terms for (b) progesterone receptor, (c) tumor size, (d) age, (e) hormonal treatment and (f) tumor grade covariates.

$1.678 tumgrad3_i + pb(age, 5.183)]$. In Figure 9, we present the estimated cured proportions for different levels of the explanatory variables as function of *age*. As suggested by a referee, we also present in this figure the estimated cure proportions for the Weibullcr model. We may note in Figure 9 (a1-b1) that the tumor grading 2 and 3 are very aggressive, influencing dramatically the cured probability. The same aggressive influence can be observed in the patients that do not receive hormonal treatment with tamoxifen. Note in Figure 9(a2-b2) that the Weibullcr model does not identify the difference between the levels of tumor grade 1 and 3. Finally, based on the results of the LSCcr model, the probability of cure increases when age increases in the range [20,45], decreases in the age range [45,60] and then stabilizes when age is greater than 60.

Figure 10 displays the estimated hazard functions. They reveal that the hazard of recurrence has a bimodal shape with high chance of failure in approximately 500 and 1500 days. We can also note in these plots the nonlinear effects of the age (see Figure 8(d)) in the hrf.

Figure 11 displays some residual plots that help verifying the adequacy and the assumptions of the selected fitted model given in (12). We also present in this figure the residual plots for the Weibullcr model. Panel (a) and (d) indicate that the normalized quantile residuals have an approximately normal distribution. Panel (e) shows that there a few points off the line in low end of the range. Further, the

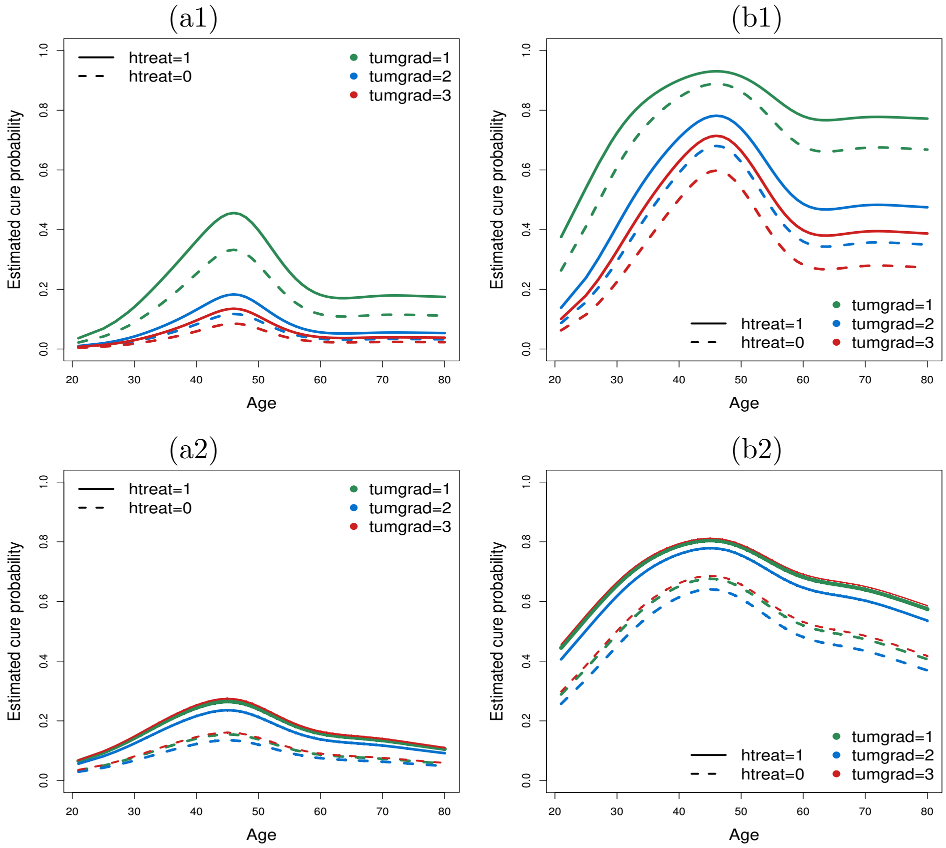

Figure 9: The estimated cured proportions for each level of *tumgrad* and *htreat* as function of *age* by taking: (a) $\min(prm) = 0$ and $tumsize = 60$ and (b) $prm = 200$ and $tumsize = 10$ for the (1) LSCcr and (2) Weibullcr models.

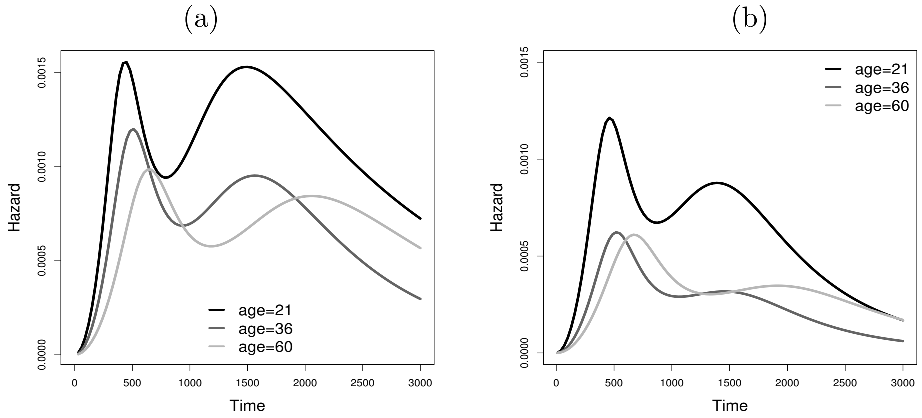

Figure 10: For the fitted LSCcr GAMLSS, the estimated hazard functions for $tumgrad = 2$, $htreat = 1$, $age = 21, 36, 60$ and considering: (a) $\min(prm) = 0$ and $tumsize = 60$ and (b) $prm = 200$ and $tumsize = 10$.

WP presented in Panel (c) indicates that there are no evidences of inadequacies on it, since all the residuals fall in the "acceptance" region inside the two elliptic curves. On the other hand, the WP presented in Panel (f) indicates failure for modelling the kurtosis. In general, the LSCcr model based on the GAMLSS framework provides a reasonable fit to these data.

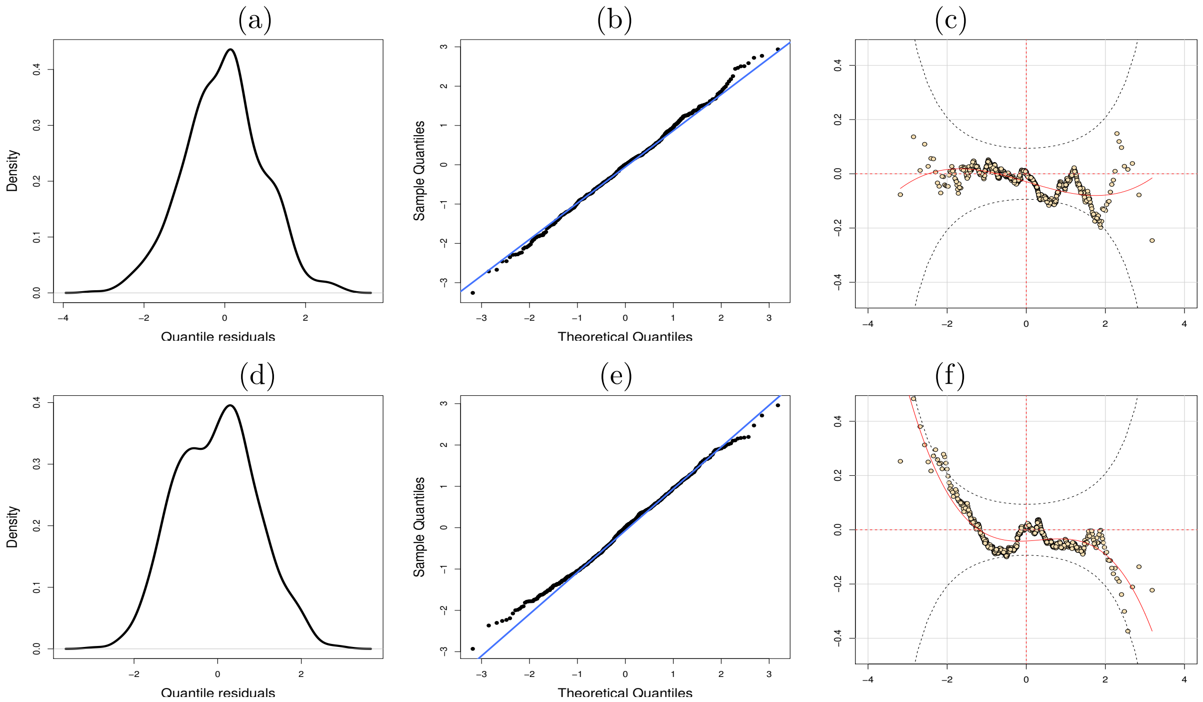

Figure 11: For the fitted LSCcr GAMLSS, (a) density of the quantile residuals, (b) Q-Q plot and (c) WP, and for the fitted Weibullcr GAMLSS, (d) density of the quantile residuals, (e) Q-Q plot and (f) WP.

6 Conclusions

The semi-parametric *log-sinh Cauchy cure rate* (LSCcr) regression model provides a flexible regression model for a dependent real outcome. The parameters of the model can be interpreted as relating to location, scale, bimodality and cure rate proportion and each of them can be modelled as parametric or smooth nonparametric functions of explanatory variables. Procedures for fitting the semi-parametric LSCcr *generalized additive model for location, scale and shape* (GAMLSS) and for model diagnostics are included in the GAMLSS package and they are available by the first author in the Appendix. We conclude by means of a study simulation that erroneous interpretations can be made if the proposed regression structure is not appropriate. A real data set is used to illustrate the usefulness of the semi-parametric LSCcr regression model, showing that it provides better performance than the usual methods in the presence of nonlinear effects in the cure rate proportion.

Acknowledgements

The first author acknowledge the financial support of the "Ciência sem Fronteiras" program of CNPq (Brazil) under the process number 200574/2015-9.

Conflict of interest

The authors have declared no conflict of interest.

Appendix: Computational codes

Here, we present the codes implemented in the `GAMLSS` package in the software `R`. The pdf, cdf, qf and the samples generator functions are

```
library(gamlss.cens); library(gamlss) #required packages
source("https://goo.gl/AppEb0") #implemented codes
dLSCc(x,mu,sigma,nu,tau) #pdf
pLSCc(x,mu,sigma,nu,tau) #cdf
qLSCc(u,mu,sigma,nu,tau) #qf
rLSCc(n,mu,sigma,nu,tau) #samples generator
```

Next, we present the codes used in the data analysis.

```
library(shrink) ;data(GBSG) ;attach(GBSG) #loading data set
#Selecting the regression model
#null model
m1=gamlss(Surv(rfst,cens) ~1, family=cens("LSCc"),c.crit=0.1, n.cyc=40)#null model
#Selecting the model for tau
m2=stepGAICAll.A(m1, scope=list(lower=~1, upper=~as.factor(htreat)+ +as.factor(tumgrad)+
                pb(age)+pb(tumsize)+ pb(prm)+ pb(esm)), mu.try = F,sigma.try = F,nu.try = F)
#Note that the effects of prm and tumsize covariates are linear.
#Now, selecting the model for mu, sigma and nu.
m3 =gamlss(Surv(rfst,cens) ~1, family=cens("LSCc"),nu.start=0.4,
    c.crit=0.01, n.cyc=40,tau.formula=~prm + tumsize+ pb(age)+ as.factor(htreat)+as.factor(
        tumgrad))
m4 =stepGAICAll.A(m3, scope=list(lower=~1,
   upper=~htreat+ as.factor(tumgrad)+ pb(age)+pb(tumsize)+ pb(prm)+ pb(esm)),
   tau.try = F,tau.start=m3$tau.fv,nu.start=0.4,n.cyc=20)
edfAll(m4);
#Note that the effects of age and  prm  covariates are linear.
#Then, the final model is
model =gamlss(Surv(rfst,cens) ~age+prm+as.factor(htreat), sigma.fo=~as.factor(tumgrad),
   nu.fo=~1,tau.fo=~prm + tumsize+ pb(age)+ as.factor(htreat)+as.factor(tumgrad),
   family=cens("LSCc"),nu.start = 0.4,c.crit=0.001, n.cyc=100)
#Diagnostic
plot( density(model$residuals),xlab="Quantile residuals",main = "",lwd=4)
qqnorm(model$residuals,pch=16); qqline (model$residuals ,col ="royalblue1",lwd=3)
wp(model)
```

References

Altman, D.G., Lausen, B., Sauerbrei, W. and Schumacher, M. (1994). Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, **86**, 829-835.

Balakrishnan, N. and Pal, S. (2012). EM algorithm-based likelihood estimation for some cure rate models. *Journal of Statistical Theory and Practice,* **6**, 698-724.

Berkson, J. and 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, Series B,* **11**, 15–53.

Buuren, S.V. and Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine,* **20**, 1259–1277.

Cancho, V.G., Dey, D.K. and Louzada, F. (2015). Unified multivariate survival model with a surviving fraction: an application to a Brazilian customer churn data. *Journal of Applied Statistics*, **43**, 572-584.

Chen, M.H., Ibrahim, J.G. and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**, 909-919.

Cooner, F., Banerjee S., Carlin, B.P. and Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association*, **102**, 560-572.

Cordeiro, G.M., Cancho, V.G., Ortega, E.M.M. and Barriga, G.D.C. (2016). A model with long-term survivors: negative binomial Birnbaum-Saunders. *Communications in Statistics - Theory and Methods*, **45**, 1370-1387.

da Cruz, J.N., Ortega, E.M.M. and Cordeiro, G.M. (2016). The log-odd log-logistic Weibull regression model: modelling, estimation, influence diagnostics and residual analysis. *Journal of Statistical Computation and Simulation*, **86**, 1516-1538.

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

Eilers, P.H. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science,* **11**, 89–121.

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

Fitzgibbons, P.L., Page, D.L., Weaver, D., Thor, A.D., Allred, D.C., Clark, G;M;, et al. (2000). Prognostic factors in breast cancer: College of American Pathologists consensus statement 1999. *Archives of pathology & laboratory medicine*, **124**, 966–978.

Gospodarowicz, M.K., O'Sullivan, B. and Sobin, L.H. (Eds.). (2006). Prognostic factors in cancer (pp. 165-168). Wiley-Liss. Frankfurt.

Green, P.J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach.* CRC Press.

Hashimoto, E.M., Ortgea, E.M.M., Cancho, V.G. and Cordeiro, G.M. (2015). A new long-term survival model with interval-censored data. *Sankhya B*, **77**, 207–239.

Hashimoto, E.M., Cordeiro, G.M., Ortega, E.M.M. and Hamedani, G.G. (2016). New flexible regression models generated by gamma random Variables with censored data. *International Journal of Statistics and Probability,* **5**, 9-31.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized additive models,* Vol. 43, CRC Press.

Ibrahim, J.G., Chen, M.H. and Sinha, D. (2001). *Bayesian Survival Analysis.* Springer: New York.

Ko, A. (2009). Everyone's guide to cancer therapy: How cancer is diagnosed, treated, and managed day to day. Andrews McMeel Publishing.

Lagakos, S.W. (1988). Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in Medicine*, **7**, 257-274.

Lanjoni, B.R., Ortega, E.M.M. and Cordeiro, G.M. (2016). Extended Burr XII regression models: Theory and applications. *Journal of Agricultural, Biological and Environmental Statistics*, **21**, 203-224.

Lee, Y., Nelder, J.A. and Pawitan, Y. *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood.* CRC Press, 2006.

Lønning, P.E. (2007). Breast cancer prognostication and prediction: are we making progress?. Annals of Oncology, 18 (suppl 8), viii3-viii7.

Maller, R.A. and Zhou, X. (1996). Survival analysis with long-term survivors. New York: Wiley.

Morgan, T.M. and Elashoff, R. M. (1986). Effect of categorizing a continuous covariate on the comparison of survival time. *Journal of the American Statistical Association*, **81**, 917-921.

Ortega, E.M.M., Cordeiro, G.M., Hashimoto, E.M. and Cooray, K. (2014). A log-linear regression model for the odd Weibull distribution with censored data. *Journal of Applied Statistics*, **41**, 1859-1880.

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

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

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

Rigby, R.A. and Stasinopoulos, D.M. (2014). Automatic smoothing parameter selection in GAMLSS with an application to centile estimation. *Statistical Methods in Medical Research,* **23**, 318–332.

Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software,* **23**, 1–46.

Schumacher, M., Bastert, G., Bojar, H., Huebner, K., *et al.* (1994). Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. *Journal of Clinical Oncology*, **12**, 2086–2093.

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

Tsodikov, A.D., Ibrahim, J.G. and Yakovlev, A.Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, **98**, 1063-1078.