**University of São Paulo
“Luiz de Queiroz” College of Agriculture**

New flexible parametric and semiparametric models for survival analysis

Thiago Gentil Ramires

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

**Piracicaba
2017**

Thiago Gentil Ramires
Degree in Statistics

New flexible parametric and semiparametric models for survival analysis

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Adviser:

Prof. Dr. **EDWIN MOISES MARCOS ORTEGA**

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2017

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Ramires, Thiago Gentil

New flexible parametric and semiparametric models for survival analysis/
Thiago Gentil Ramires. -- versão revisada de acordo com a resolução CoPGr
6018 de 2011. -- Piracicaba, 2017 .
128 p.

Tese (Doutorado) -- USP / Escola Superior de Agricultura "Luiz de
Queiroz".

1. Bimodalidade 2. GAMLSS 3. P-splines 4. Fração de cura . I. Título.

DEDICATION

To my parents,
Ademir Ramires and **Janet Gentil Ramires**, for all the love and dedication they have for me.

To my girlfriend,
Ana Julia Righetto who guided my ways here.

To my brother,
Juliano Gentil Ramires that even being distant, he always remembers what it means to be a brother.

To them,
I lovingly dedicate this work.

ACKNOWLEDGMENTS

Agradeço primeiramente a meus pais e irmãos, por me apoiar sempre nas minhas escolhas e conquistas dos meus sonhos.

À meu amigo e orientador Edwin Moises Marcos Ortega, o qual me incentivou e colaborou a conquistar mais essa conquista em minha vida. Muito obrigado por tudo. Também ao Prof. Gauss Cordeiro, o qual serei eternamente grato por todas as considerações e motivações que me proporcionou até o momento.

To my advisor in Belgium, Niel Hens, which gave me all the support during the PhD sandwich period.

À minha namorada Ana Julia Righetto, a qual sem ela, não teria conseguido chegar até aqui. Obrigado por tudo, por ter ficado ao meu lado nos momentos mais difíceis da minha vida. Serei eternamente grato.

À todos amigos de Piracicaba que sempre estiveram ao meu lado nos melhores e piores momentos em especial: Rodrigo Pescim, Pedro Cerqueira, Guilherme Biz, Lucas Santana, Luiz Ricardo Nakamura, Djair Durand, Thiago Oliveira, Alexandre Lavorenti, Renan Pinto, Rafael Jacomini, André Sanches, Henrique Gioia e Pedro Lian Barbieri.

To my friends in Belgium, Shah Rukh Sajid, Svitlana Railian, Flávio Rabelo, Lucélia Borgo, Sain Lordgilani, Sarmad Zaman and Tooba Moosa. Thank you guys, feel very important in my life.

À meus melhores amigos Luiz Fernando Navarro, Fabio Antonietti, Pedro Henrique Baggio, Gabriel Polizel Santos, Fábio Casagrande Basseto, Gustavo Gomes Correia, Jeanmichel Cavalaro, que mesmo distante mostram que uma amizade verdadeira é para sempre.

À Luciane Brajão, Solange Sabadin e Rosni Pinto que nesses anos se tornaram partes essenciais de nossas vidas.

À CNPq - pela bolsa de mestrado concedida no doutorado e doutorado sanduíche.

À todos os professores que convivi durante o curso de mestrado e doutorado em Estatística e Experimentação Agronômica, me dando a oportunidade de participar desta família ESALQ, fornecendo conhecimentos e possibilidades de alto nível, com as quais tive a oportunidade de trabalhar em minhas pesquisas.

Aos alunos do curso de Pós-Graduação em Estatística e Experimentação Agronômica da ESALQ/USP, os quais fizeram parte desta fase.

Enfim, a todos os amigos que me ajudaram a compor mais um pedaço da minha história.

CONTENTS

Resumo	8
Abstract	9
1 Introduction	11
References	12
2 A bimodal flexible distribution for lifetime data	15
2.1 Introduction	15
2.2 The ELSC model	16
2.3 Expansion of the quantile function	18
2.4 Moments	20
2.5 Other measures	21
2.5.1 Generating function	21
2.5.2 Mean deviations	22
2.5.3 Order statistics	22
2.6 Inference	22
2.7 Simulation	23
2.8 Applications	26
2.8.1 Eruption data	26
2.8.2 Efron data	27
2.8.3 Entomology data	29
2.9 Program description	30
2.10 Conclusions	31
References	31
3 New regression model with four regression structures and computational aspects	33
3.1 Introduction	33
3.2 Properties of the standardized ESC distribution	34
3.2.1 Expansion of the quantile function	35
3.2.2 Moments	36
3.3 The ESC regression model	36
3.3.1 Definition	37
3.3.2 Estimation	38
3.4 Simulation Study	39
3.4.1 Location simulation	39
3.4.2 GAMLSS simulation	40
3.5 Study of model misspecification	41
3.6 Sensitivity and residual analysis	42
3.6.1 Global influence	42
3.6.2 Local influence	42
3.6.2.1 Case-weight perturbation	43
3.6.2.2 Response perturbation	43
3.6.2.3 Explanatory variable perturbation	43
3.6.3 Residual Analysis	44
3.7 Applications	45
3.7.1 Shrimp data	45
3.7.1.1 Global influence analysis	46
3.7.1.2 Local influence analysis	46

3.7.1.3	Residual analysis	47
3.7.2	Entomology data	48
3.7.2.1	Global influence analysis	49
3.7.2.2	Local influence analysis	49
3.7.2.3	Residual analysis	50
3.8	Conclusions	51
3.9	Script for the ESC regression model	52
References		52
4	A flexible bimodal model with long-term survivors and different regression structures	55
4.1	Introduction	55
4.2	The ELSC model for survival data with long-term survivors	56
4.2.1	Definition	57
4.3	Regression model	57
4.3.1	Parametric model	58
4.3.2	Related models	59
4.3.3	Inference	59
4.3.4	Selecting explanatory variables and link functions	60
4.4	Goodness of fit, diagnostics and influence measures	60
4.4.1	Choosing the best model	61
4.4.2	Diagnostic and influence analysis	61
4.5	Simulation	62
4.5.1	Simulation 1: ELSCcr model	62
4.5.2	Simulation 2: ELSCcr regression model	63
4.6	Applications	64
4.6.1	Calving data	64
4.6.2	Gastric cancer data	65
4.6.3	Breast cancer data	67
References		68
4.7	Conclusions	69
5	Predicting the cure rate of breast cancer using a new regression model with four regression structures	73
5.1	Introduction	73
5.2	The LSCp model	75
5.3	Regression models	77
5.3.1	Definition	77
5.3.2	Inference	77
5.4	Model selection	78
5.4.1	Select the distribution	79
5.4.2	Selecting explanatory variables	79
5.4.3	Diagnostics	79
5.4.4	Global influence	80
5.5	Simulation study	80
5.6	Predicting breast cancer data	82
5.7	Conclusions	86
5.8	Supplementary material	87
5.8.1	Codes used in global influence	87
5.8.2	Codes used in simulation study	88

5.8.3 Codes of the Weibullcr GAMLSS	89
References	89
6 A flexible semiparametric regression model for bimodal, asymmetric and censored data	91
6.1 Introduction	91
6.2 The ESC regression model	92
6.2.1 Definition	92
6.2.2 Nonparametric additive functions	93
6.2.3 Estimation	94
6.2.4 Model strategy	95
6.2.5 Simulation	96
6.2.6 Diagnostics	96
6.2.7 Global influence	97
6.3 Simulation Study	97
6.4 Applications	99
6.4.1 Application: Body mass data	99
6.5 Eruption data	101
6.6 Conclusions	106
References	106
7 Estimating nonlinear effects in regression models with long-term survivors	109
7.1 Introduction	109
7.2 The Log sinh Cauchy GAMLSS with long-term survivors	110
7.2.1 The LSCcr distribution	111
7.3 The LSCcr GAMLSS	112
7.4 Model selection	114
7.4.1 Inference	114
7.4.2 Goodness-of-fit	115
7.4.3 Additive terms selection	115
7.5 Simulation study	116
7.6 Predicting the cure rate of breast cancer	117
7.7 Conclusions	121
References	122
8 Conclusion	125
APPENDICES	127

RESUMO

Novos modelos flexíveis paramétricos e semi-paramétricos para análise de sobrevivência

Nesse trabalho foi proposto uma nova distribuição, denominada de exponentiated log-sinh Cauchy, a qual possui densidades bimodais e pode ser utilizada como alternativa aos modelos de mistura. Com base na nova distribuição, foram propostos: modelos de regressão baseados nos modelos GAMLSS; modelos com fração de cura baseados em modelos de mistura e tempo de promoção; modelo semi-paramétrico modelando os parâmetros com splines penalizados; modelo semi-paramétrico com fração de cura utilizando splines para modelar efeitos não lineares na proporção de curados. Para todos os modelos propostos, toda parte computacional foi implementada no *software* R, sendo disponibilizada ao longo do documento assim como breve descrições de uso.

Palavras-chave: Bimodalidade, GAMLSS, P-splines, Fração de cura

ABSTRACT

New flexible parametric and semiparametric models for survival analysis

In this work was proposed a new distributions, called log-sinh Cauchy, with has bimodal shapes and can be used as alternative to the mixture models. Based in the proposed distribution, the following models were proposed: Regression model based in the GAMLSS framework; models with cure rate based in the mixture and promotion time models; semiparametric models, modeling the parameters using penalized splines; semiparametric models, using the penalized splines to model the non-linear effects present in the cure rate. For all proposed models, the computational codes were implemented in the R software, with is available along of the document as well as some brief introduction on how to use them.

Keywords: Bimodality, GAMLSS, P-splines, Cure rate

1 INTRODUCTION

Present in virtually all areas, statistics is an outstanding tool for data analysis. Among these areas is the survival analysis, which has applications in several areas of research, like medicine, agronomy, engineering, biology, economics and other areas related to health and finance. The increasing use of statistics is due in part to the development of more efficient techniques and methods along with technological and computational advances that allow the creation of more sophisticated models for analyzing data with different behavior than usually found in case studies in the literature.

With the ease of database construction, new density behaviors related to variable responses are emerging, which in some situations require extremely complicated shapes and more complex models. Recently, several models have been proposed in the survival analysis literature, which have greater flexibility, resulting in more accurate estimates and analyses. Mixtures and transformations between distributions generate interesting results when applying probability density failure or risk rate functions. Over the past 10 years, hundreds of new models have been proposed in the survival analysis literature, for which a brief discussion can be found in Tahir and Nadarajah (2015).

Among the different proposed models, it is notable that only a small number take bimodal forms. Data that exhibit bimodal behavior arise in many different disciplines. In medicine, urine mercury excretion has two peaks, see for example, Ely *et al.* (1999). In material characterization, in a study conducted by Dierickx *et al.* (2000), grain size distribution data revealed a bimodal structure. In meteorology, Zhang *et al.* (2003) found that water vapor levels in tropical regions commonly have bimodal distributions. Furthermore, most models that are able to assume bimodal forms have positive skewness, and are inefficient to fit symmetric or negative symmetry. Alternatively, many authors have used mixture distributions to model data with bimodal behavior, associating a specific distribution for each modal region.

Due to the need for new models capable of capturing bimodal forms, present a new model for survival analysis called “exponentiated log-sinh Cauchy”, which has four parameters and is able to take symmetrical bimodal or positive or negative asymmetrical shape. Properties, applications, simulations and computational implementation of the new model are also presented.

In many cases, the response variable’s behavior is influenced by other variables, called explanatory variables. In such cases, it is necessary to add these variables in statistical models to achieve better interpretation. One of the most common methods of relating the explanatory variables with the response variable is to use the class of location-scale models. But location-scale models relate only the location parameter with the explanatory variables, so in many cases it is necessary to use more complex models to get a good fit, which would not be needed if the scaling, kurtosis or others parameters were also modeled by explanatory variables. In this sense, we present a regression model, based in the exponentiated log-sinh Cauchy model, which belongs to the “generalized additive models for location, scale and shape” (GAMLSS) class of models (Rigby and Stasinopoulos, 2005).

The advantage of the class when compared to the location-scale class of models is that all parameters can be explained by explanatory variables, which in the case of exponentiated log-sinh Cauchy model are the location, scale, bimodality and skewness parameters. All computational scripts of the new regression model were implemented in the R software (R Core Team, 2015) using the GAMLSS package (Stasinopoulos and Rigby, 2007) and are available, for easy use by anyone familiar with the R software.

Models for survival analysis typically consider that every subject in the study population is susceptible to the event under study and will eventually experience such event if follow-up is sufficiently long. However, there are situations when a fraction of individuals are not expected to experience the event of interest, that is, those individuals are cured or not susceptible. Based in the mixture models (MMs) pioneered by Boag (1949), Berkson and Gage (1952), we propose a new cure rate model based

on the “exponentiated log-sinh Cauchy” distribution. Using the GAMLSS framework, we can model the location, scale, bimodality, skewness and cure rate parameters. Based on the promotion time cure models (Yakovlev and Tsodikov, 1996), we also proposed a new model to estimate breast carcinoma mortality, assuming that the number of competing causes that can influence the survival time follows a Poisson distribution.

When using the parametric regression models belonging to the class of location-scale or GAMLSS models, in many situations the explanatory variables do not have a linear relation with the dependent variable, requiring the use of nonlinear functions to explain its behavior. Among various nonlinear functions, the splines (the focus of this paper) stand out for being extremely flexible in capturing various types of behavior. Currently splines are used especially considering the Cox models (Cox, 1972). Although becoming more popular in the literature, there are few references on the use of splines in the class of location-scale and GMLSS models.

In this context, we propose a new semiparametric heteroscedastic regression model allowing for positive and negative skewness and bimodal shapes using the B-spline basis for nonlinear effects. The proposed distribution is based on the generalized additive models for location, scale and shape framework in order to model any or all the parameters of the distribution using parametric linear and/or nonparametric smooth functions of explanatory variables. Finally the idea of the semiparametric models are extended for the new cure rate models, being possible to estimate nonlinear effects of explanatory variable in the cure rate parameter.

References

- 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.
- Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187–220.
- Dierickx, D., Basu, B., Vleugels, J. and Van der Biest, O. (2000). Statistical extreme value modeling of particle size distributions: experimental grain size distribution type estimation and parameterization of sintered zirconia. *Materials characterization*, **45**, 61–70.
- Ely, J.T.A., Fudenberg, H.H., Muirhead, R.J., LaMarche, M.G., Krone, C.A., Buscher, D. and Stern, E.A. (1999). Urine mercury in micromercurialism: bimodal distribution and diagnostic implications. *Bulletin of environmental contamination and toxicology*, **63**, 553–559.
- 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.
- 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.
- Tahir, M.H. and Nadarajah, S. (2015). Parameter induction in continuous univariate distributions: Well-established G families. *Anais da Academia Brasileira de Ciências*, **87**, 539–568.
- Team, R.C. (2000). R Language Definition.
- Zhang, C., Mapes, B.E. and Soden, B. J. (2003). Bimodality in tropical water vapour. *Quarterly Journal of the Royal Meteorological Society*, **129**, 2847–2866.

Yakovlev A and Tsodikov AD. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications. Mathematical Biology and Medicine, Vol. 1.* World Scientific, New Jersey.

2 A BIMODAL FLEXIBLE DISTRIBUTION FOR LIFETIME DATA

Abstract: A four-parameter extended bimodal lifetime model called the exponentiated log-sinh Cauchy distribution is proposed. It extends the log-sinh Cauchy and folded Cauchy distributions. We derive some of its mathematical properties including explicit expressions for the ordinary moments and generating and quantile functions. The method of maximum likelihood is used to estimate the model parameters. We implement the fit of the model in the GAMLSS package and provide the codes. The flexibility of the model is illustrated by means of three real data sets.

Keywords: Bi-modality; Exponentiated sinh Cauchy distribution; GAMLSS; Lifetime distribution.

2.1 Introduction

Generalizing lifetime distributions by introducing a few extra shape parameters is an essential method to better explore the skewness and the tails and other properties of the transformed distributions. Following the latest trend, applied statisticians are now able to construct more generalized distributions, which provide better goodness-of-fit measures when fitted to real data rather than by using the classical distributions. The Weibull, log-normal and log-logistic are very popular distributions for modeling lifetime data and phenomenon with unimodal and monotone failure rates. In these cases, they may be chosen because of their negatively and positively skewed density shapes. However, these models do not provide reasonable parametric fits for modeling phenomenon with non-monotone failure rates such as the bathtub shaped and bimodal failure rates, which are common in reliability and biological studies. In this paper, we study a four-parameter generalization of the exponentiated sinh Cauchy (ESC) distribution on the basis of the sinh Cauchy (SC) model, both proposed by Cooray (2013), for modeling bimodal and unimodal data. The advantage of this approach for constructing a parametric family of distributions lies in its flexibility to model both bathtub and bimodal failure rates even though the baseline failure rate may be monotonic. The generated model is called the *exponentiated log-sinh Cauchy* (ELSC) distribution. As we will see later, its hazard rate function (hrf) can be constant, decreasing, increasing, upside-down bathtub (unimodal), bathtub and bimodal shaped. Due to the great flexibility of the ELSC hrf, it thus provides a good alternative to many existing life distributions in modeling positive real data sets.

Cooray (2013) applied the hyperbolic sine transformation to the standard Cauchy distribution by defining the SC model, whose cumulative density function (cdf) is given by

$$\Pi(y) = \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh \left(\frac{y - \mu}{\sigma} \right) \right], \quad y \in R, \quad (2.1)$$

where $\mu \in R$ 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 SC distribution produces both bimodal and unimodal densities with a wide range of tail weights. It has a real support and therefore is not appropriate for survival data. As a better alternative, we present the *log-sinh Cauchy* (LSC) model.

Let Y be a random variable having cdf (2.1). The random variable $X = e^Y$ defines the LSC distribution, whose cdf is given by

$$G(x) = \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh \left(\frac{\log(x) - \mu}{\sigma} \right) \right], \quad x > 0. \quad (2.2)$$

The SC and LSC models are not appropriate for modeling real data, even though they have some theoretical advantages due to their symmetric nature. To provide an asymmetry for the SC distribution, Cooray (2013) proposed the ESC distribution using the exponentiated class of distributions (Gupta and Kundu, 2001). The cdf of the exponentiated class is given by

$$F(x) = G(x)^\tau, \quad (2.3)$$

where $G(x)$ is the parent cdf and $\tau > 0$ denotes an extra power shape parameter. By differentiating (2.3), the probability density function (pdf) of the exponentiated class is given by

$$f(x) = \tau G(x)^{\tau-1} g(x), \quad (2.4)$$

where $g(x)$ is the baseline pdf.

The paper is outlined as follows. In Section 2.2, we define the ELSC model by applying the exponentiated generator to the LSC distribution. In Section 2.3, we derive a power series for the quantile function (qf) of this distribution. In Section 2.4, we obtain explicit expressions for its moments. A range of its mathematical properties is explored in Section 2.5 including generating function, mean deviations and order statistics. The estimation of the model parameters by maximum likelihood is addressed in Section 2.6. The performance of the maximum likelihood estimators (MLEs) is investigated through a simulation study in Section 2.7. Applications to three real data sets are addressed in Section 2.8 to prove empirically the flexibility of the model. In Section 2.9, we provide a brief discussion of the template for the ELSC distribution implemented in the ‘‘GAMLSS’’ R package (Stasinopoulos and Rigby, 2007). We also provide the computational codes used in the applications. Finally, Section 2.10 ends with some conclusions.

2.2 The ELSC model

We can add skewness for an extended LSC distribution by adopting the exponentiated class of distributions (Gupta and Kundu, 2001) given by (2.3). Inserting (2.2) in equation (2.3), the ELSC cdf is given by

$$F(x; \mu, \sigma, \nu, \tau) = \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan [\nu \sinh(w)] \right\}^{\tau}, \quad (2.5)$$

where $w = [\log(x) - \mu]/\sigma$. For $\tau = 1$, the LSC distribution is just a special case of (2.5). The pdf corresponding to (2.5) is given by

$$f(x; \mu, \sigma, \nu, \tau) = \frac{\tau \nu}{x \sigma \pi} \frac{\cosh(w)}{[\nu^2 \sinh^2(w) + 1]} \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan [\nu \sinh(w)] \right\}^{\tau-1}. \quad (2.6)$$

Henceforth, let $X \sim \text{ELSC}(\mu, \sigma, \nu, \tau)$ be a random variable with density function (2.6). We can omit sometimes the dependence on the parameters and write simply $f(x) = f(x; \mu, \sigma, \nu, \tau)$.

The survival function and hrf of X are given by $S(x) = 1 - F(x)$ and $h(x) = f(x)/S(x)$, respectively. Plots of the ELSC density, survival and hazard functions for selected parameter values are displayed in Figures 2.1, 2.2 and 2.3, respectively.

In Figure 2.1a-b, we check the effects of the location and scale parameters μ and σ on the function $f(x)$. Figure 2.1c reveals clearly the bi-modality effect caused by the parameter ν . Further, Figure 2.1d reveals that the density of X is bimodal and symmetric, bimodal and right-skewed, bimodal and left-skewed depending on the parameter τ . Figures 2.3a and 2.3b indicate that the hrf of X has decreasing, unimodal and bimodal forms and double bathtub-shaped and unimodal and bathtub-shaped, respectively.

We provide in Figures 2.4a-b a numerical investigation to identify how the parameter values change the shapes of the hrf of X for some parameter ranges. Based on these plots, we can obtain bimodal shapes for the hrf of X for small values of the parameters ν and τ . However, large values of these parameters are necessary to obtain this characteristic when the parameter σ increases.

Because of the current computational facilities, several researchers construct new lifetime models to facilitate their use in lifetime data analysis. It is a common practical technique to fit new models to real data and develop scripts in statistical software R (R Core Team, 2015). DeCastro *et al.* (2010)

Figure 2.1. Plots of the ELSC density for fixed values of: (a) $\sigma = 0.1$, $\nu = 0.2$ and $\tau = 1$; (b) $\mu = 4$, $\nu = 0.3$ and $\tau = 0.7$; (c) $\mu = 4$, $\sigma = 0.1$ and $\tau = 1$; (d) $\mu = 4$, $\sigma = 0.1$ and $\nu = 0.2$.

Figure 2.2. The ELSC survival function when $\mu = 4$, $\sigma = 0.1$ and: (a) For $\tau = 1$ and different values of ν ; (b) For $\nu = 0.05$ and different values of τ .

implemented some long-term survival models by taking the Weibull as the parent distribution. Rodrigues *et al.* (2009) implemented the COM–Poisson cure rate model and illustrate its flexibility by means of a real data set. Following these ideas, the ELSC model is implemented in the R software, where a short discussion is given in Section 2.9.

Figure 2.3. The ELSC hrf: (a) For $\tau = 1$ and different values of μ , σ and ν ; (b) For $\mu = 4$ and $\tau = 0.01$ and different values of σ and ν .

Figure 2.4. The ELSC hrf shapes as functions of ν and τ for $\mu = 1$ and: (a) $\sigma = 0.4$; (b) $\sigma = 0.7$.

2.3 Expansion of the quantile function

Inverting $F(x) = u$ (for $0 < u < 1$), we obtain the qf of X

$$x = Q(u) = \exp \left(\mu + \sigma \operatorname{arcsinh} \left\{ \frac{1}{\nu} \tan \left[\pi \left(u^{1/\tau} - 0.5 \right) \right] \right\} \right). \quad (2.7)$$

Quantiles of interest can be obtained from (2.7) by substituting appropriate values for u . In particular, the median of X is obtained when $u = 1/2$. We can also use (2.7) for simulating ELSC random variables by setting u as a uniform random variable in the unit interval $(0, 1)$. The qf of the LSC distribution can be obtained by taking $\tau = 1$ in equation (2.7).

Next, we derive an expansion for the qf of X to obtain some ELSC properties in the following sections. Expanding (2.7) in power series using *Mathematica*, we obtain

$$Q(u) = e^\mu \exp \left(\sum_{k=0}^{\infty} c_k z^{2k+1} \right),$$

where $z = u^{1/\tau} - 0.5$, $c_k = \frac{\sigma b_k}{(2k+1)!} \left(\frac{\pi}{\nu} \right)^{2k+1}$ and $b_0 = 1$, $b_1 = (2\nu^2 - 1)$, $b_2 = (16\nu^4 - 20\nu^2 + 9)$, $b_3 = (272\nu^6 - 616\nu^4 + 630\nu^2 - 225)$, $b_4 = (7936\nu^8 - 28160\nu^6 + 48384\nu^4 - 37800\nu^2 + 11025)$, ...

By simple transformation of quantities, we can write

$$Q(u) = e^\mu \exp \left(\sum_{k=1}^{\infty} \frac{d_k}{k!} z^k \right), \quad (2.8)$$

where

$$d_{2j} = 0 \quad \text{for } j = 1, 2, \dots \quad \text{and} \quad d_{2j+1} = (2j+1)! c_j \quad \text{for } j = 0, 1, 2, \dots \quad (2.9)$$

We can use the Bell polynomials¹ to rewrite equation (2.8). The exponential partial Bell polynomials in formal double series expansion are defined by Comtet (1974, p.133) as

$$\exp \left(u \sum_{m \geq 1} x_m \frac{t^m}{m!} \right) = \sum_{n,k \geq 0} \frac{B_{n,k}}{n!} t^n u^k, \quad (2.10)$$

where

$$B_{n,k} = B_{n,k}(x_1, x_2, \dots, x_{n-k+1}) = \sum \frac{n!}{c_1! c_2! \dots (1!)^{c_1} (2!)^{c_2} \dots} x_1^{c_1} x_2^{c_2} \dots,$$

and the summation is over all integers $c_1, c_2, c_3, \dots \geq 0$ such that $c_1 + 2c_2 + 3c_3 + \dots = n$ and $c_1 + c_2 + c_3 + \dots = k$. These exponential partial Bell polynomials can be evaluated in **Mathematica** and **Maple** using **BellY[n,k,{x₁,...,x_{n-k+1}}]** and

IncompleteBellB(n, k, x[1], z[2], ..., x[n-k+1]).

Using the definition of the complete Bell polynomials and (2.10), equation (2.8) can be expressed as

$$Q(u) = e^\mu \sum_{k=0}^{\infty} \frac{B_k(d_1, \dots, d_k)}{k!} z^k,$$

where $B_k = B_k(d_1, \dots, d_k) = \sum_{r=1}^k B_{k,r}(d_1, \dots, d_{k-r+1})$ (for $k \geq 0$) is the complete Bell polynomial of order k .

The coefficients B_k can be easily obtained using **Mathematica**, **Maple** and **Sage** softwares. Replacing z in the last equation, the qf of X can be rewritten as

$$Q(u) = e^\mu \sum_{k=0}^{\infty} \frac{B_k(d_1, \dots, d_k)}{k!} (u^{1/\tau} - 0.5)^k. \quad (2.11)$$

By expanding the binomial term, we have

$$Q(u) = e^\mu \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \frac{(-1)^{k-j}}{2^{k-j} k!} \binom{k}{j} B_k(d_1, \dots, d_k).$$

Further, changing $\sum_{k=0}^{\infty} \sum_{j=0}^{\infty}$ by $\sum_{j=0}^{\infty} \sum_{k=j}^{\infty}$, we can write

$$Q(u) = \sum_{j=0}^{\infty} p_j u^{j/\tau}, \quad (2.12)$$

where the coefficients

$$p_j = e^\mu \sum_{k=j}^{\infty} \frac{(-1)^{k-j}}{2^{k-j} k!} \binom{k}{j} B_k(d_1, \dots, d_k) \quad (2.13)$$

can be evaluated using the analytical softwares cited before.

Let $W(\cdot)$ be any integrable function in the positive real line. We can write from (2.6) and (2.12)

$$\int_0^\infty W(x) f(x; \mu, \sigma, \nu, \tau) dx = \int_0^1 W \left(\sum_{j=0}^{\infty} p_j u^{j/\tau} \right) du. \quad (2.14)$$

Equation (2.14) is an important result since it allows to obtain various mathematical properties for the ELSC distribution using integrals over $(0, 1)$. For the great majority of the applications of (2.14),

¹http://en.wikipedia.org/wiki/Bell_polynomials

we can adopt ten terms in the power series. Equations (2.12) and (2.14) are the main results of this section. The formulae derived throughout the paper can be easily handled in most symbolic computation software platforms such as those cited before. They have currently the ability to deal with analytic expressions of formidable size and complexity. Established explicit expressions to evaluate statistical measures can be more efficient than computing them directly by numerical integration.

2.4 Moments

Some of the most important features and characteristics of a distribution can be studied through moments (e.g., tendency, dispersion, skewness and kurtosis). Using (2.4), the n th moment of X can be expressed as

$$\mu'_n = E(X^n) = \tau \int_0^\infty x^n G(x)^{\tau-1} g(x) dx = \tau \int_0^1 Q_{\text{LSC}}(u)^n u^{\tau-1} du, \quad (2.15)$$

where $Q_{\text{LSC}}(u)$ denotes the qf of the LSC distribution.

Here, we give two explicit expressions for μ'_n . For the first one, we use the power series for $Q_{\text{LSC}}(u)^n$, which follows by changing μ by $n\mu$, σ by $n\sigma$ and taking $\tau = 1$ in (2.11). We have

$$Q_{\text{LSC}}(u)^n = e^{n\mu} \sum_{k=0}^{\infty} \frac{B_k(d_1^*, \dots, d_k^*)}{k!} (u - 0.5)^k, \quad (2.16)$$

where

$$d_{2j}^* = 0 \quad \text{for } j = 1, 2, \dots, \quad d_{2j+1}^* = (2j+1)! c_j^* \quad \text{for } j = 0, 1, 2, \dots \quad (2.17)$$

and $c_k^* = k \sigma b_k \pi^{2k+1} / (2k+1)!$.

Replacing (2.16) in equation (2.15), we have

$$\mu'_n = \tau e^{n\mu} \sum_{k=0}^{\infty} \frac{B_k(d_1^*, \dots, d_k^*)}{k!} \int_0^1 (u - 0.5)^k u^{\tau-1} du.$$

Let ${}_2F_1(p, q; r; y) = \sum_{j=0}^{\infty} (p)_j (q)_j y^j / [(r)_j j!]$ be the hypergeometric function, $(p)_j$ the Pochhammer symbol defined by $(p)_j = p(p+1) \cdots (p+j-1) = \Gamma(p+j)/\Gamma(p) = (-1)^j \Gamma(1-p)/\Gamma(1-p-j)$, and $\Gamma(\cdot)$ the gamma function.

The last equation can be expressed in terms of the hypergeometric function² as

$$\mu'_n = e^{n\mu} \sum_{k=0}^{\infty} \frac{(-1)^k}{2^k k!} {}_2F_1(-k, \tau; \tau+1; 2) B_k(d_1^*, \dots, d_k^*). \quad (2.18)$$

The hypergeometric function ${}_2F_1(p, q; r; y)$ can be evaluated from **Mathematica** and **Maple** as **HypergeometricPFQ**[{p,q},{r},y] and **Hypergeometric**([p,q],[r],y), respectively.

The second expression for μ'_n can be determined using (2.7) and (2.12) in equation (2.15) and changing μ by $n\mu$, σ by $n\sigma$ and setting $\tau = 1$. We obtain

$$\mu'_n = \tau \sum_{j=0}^{\infty} \frac{p_j^*}{j + \tau}, \quad (2.19)$$

where $p_j^* = e^{n\mu} \sum_{k=j}^{\infty} \frac{(-1)^{k-j}}{2^{k-j} k!} \binom{k}{j} B_k(d_1^*, \dots, d_k^*)$ and d_k^* is defined by (2.17).

Equations (2.18) and (2.19) are the main results of this section. The central moments (μ_s) and cumulants (κ_s) of X are determined as $\mu_s = \sum_{k=0}^s (-1)^k \binom{s}{k} \mu_1'^s \mu_{s-k}'$ and $\kappa_s = \mu_s' - \sum_{k=1}^{s-1} \binom{s-1}{k-1} \kappa_k \mu_{s-k}'$,

²<http://mathworld.wolfram.com/HypergeometricFunction.html>

respectively, where $\kappa_1 = \mu'_1$. The skewness $\gamma_1 = \kappa_3/\kappa_2^{3/2}$ and kurtosis $\gamma_2 = \kappa_4/\kappa_2^2$ follow from the third and fourth standardized cumulants, respectively.

When these moments do not exist, for example, for the Cauchy, Lvy and Pareto distributions, alternative measures for the skewness and kurtosis, based on qfs, are sometimes more appropriate for these distributions. The measures of skewness $\mathcal{B}$ (Galton, 1883) and kurtosis $\mathcal{M}$ (Moors, 1988) are given by

$$\mathcal{B} = \frac{Q(6/8) + Q(2/8) - 2Q(4/8)}{Q(6/8) - Q(2/8)} \quad \text{and} \quad \mathcal{M} = \frac{Q(7/8) - Q(5/8) + Q(3/8) - Q(1/8)}{Q(6/8) - Q(2/8)},$$

respectively.

For the ELSC and LSC distributions, Galton's skewness and Moors' kurtosis can be computed using the qf (2.7). Figure 2.5 displays some plots of the measures $\mathcal{B}$ and $\mathcal{M}$ as functions of the shape and bi-modality parameters. The additional shape parameter τ has substantial effect on the skewness and kurtosis of X .

Figure 2.5. Plots of the measures (a) $\mathcal{B}$ and (b) $\mathcal{M}$ as functions of τ and ν for $\mu = 3$ and $\sigma = 0.2$.

2.5 Other measures

In this section, we derive the generating function, mean deviations and order statistics of X .

2.5.1 Generating function

The moment generating function (mgf) $M(t) = E(e^{tX})$ of X can be determined from equation (2.4) in terms of its qf. We have

$$M(t) = \tau \int_0^\infty e^{tx} G(x)^{\tau-1} g(x) dx = \tau \int_0^1 u^{\tau-1} \exp[t Q_{\text{LSC}}(u)] du.$$

Combining equations (2.8) and (2.12) when $\tau = 1$, the mgf of X can be written as

$$M(t) = \tau e^{tp_0} \int_0^1 u^{\tau-1} \exp\left(\sum_{j=1}^\infty \frac{p_j^{**} u^j}{j!}\right) du,$$

where $p_j^{**} = t p_j j!$ and p_j is given by (2.13). Using again the complete Bell polynomials, we have

$$\exp\left(\sum_{j=1}^\infty \frac{p_j^{**} u^j}{j!}\right) = \sum_{j=0}^\infty \frac{B_j(p_1^{**}, \dots, p_j^{**})}{j!} u^j,$$

and then, the mgf of X follows as

$$M(t) = \tau e^{tp_0} \sum_{j=0}^\infty \frac{B_j(p_1^{**}, \dots, p_j^{**})}{(\tau + j) j!}.$$

2.5.2 Mean deviations

For empirical purposes, the first incomplete moment $m_1(s) = \int_{-\infty}^s x f(x) dx$ plays an important role for measuring inequality, for example, mean deviations and Lorenz and Bonferroni curves. A formula for $m_1(s)$ follows by setting $u = G(x)$ in (2.4) as

$$m_1(s) = \tau \int_0^s Q_{\text{LSC}}(u) u^{\tau-1} du. \quad (2.20)$$

Here, we provide two alternatives to compute the first incomplete moment of X . First, $m_1(s)$ can be derived from (2.18) by taking $n = 1$ as

$$m_1(s) = \tau e^\mu \sum_{k=0}^{\infty} \frac{(1-2s)^{-k} (s-0.5)^k s^\tau}{\tau k!} {}_2F_1(-k, \tau; \tau+1; 2s) B_k(d_1, \dots, d_k), \quad (2.21)$$

where d_k is given by (2.9). A second formula for $m_1(s)$ can be derived by inserting (2.12) in equation (2.20) and setting $\tau = 1$ as

$$m_1(s) = \tau \sum_{j=0}^{\infty} p_j \frac{s^{\tau+j}}{\tau+j}. \quad (2.22)$$

The main applications of equations (2.21) or (2.22) are related to the Bonferroni and Lorenz curves defined (for a given probability π) by $B(\pi) = m_1(q)/(\pi\mu'_1)$ and $L(\pi) = m_1(q)/\mu'_1$, respectively, where $\mu'_1 = E(X)$ and $q = Q(\pi)$ is the qf of X at π obtained from (2.7).

The mean deviations about the mean ($\delta_1 = E(|X - \mu'_1|)$) and the median ($\delta_2 = E(|X - M|)$) of X are given by

$$\delta_1(X) = 2\mu'_1 F(\mu'_1) - 2m_1(\mu'_1) \quad \text{and} \quad \delta_2(X) = \mu'_1 - 2m_1(M), \quad (2.23)$$

respectively, where $M = \text{Median}(X) = Q(0.5)$ is the median, $F(\mu'_1)$ is easily evaluated from the cdf (2.5) and $m_1(z)$ is given by (2.21) or (2.22).

2.5.3 Order statistics

Order statistics make their appearance in many areas of statistical theory and practice. Suppose $X_1, \dots, X_n$ is a random sample from the ELSC distribution. Let $X_{i:n}$ denote the i th order statistic. Using (2.5) and (2.6), the pdf of $X_{i:n}$ can be expressed as

$$\begin{aligned} f_{i:n}(x) &= K f(x) F(x)^{i-1} \{1 - F(x)\}^{n-i} = K \sum_{j=0}^{n-i} (-1)^j \binom{n-i}{j} f(x) F(x)^{j+i-1} \\ &= K \sum_{j=0}^{n-i} (-1)^j \binom{n-i}{j} \frac{\tau\nu}{x\sigma\pi} \frac{\cosh(w)}{[\nu^2 \sinh^2(w) + 1]} \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu \sinh(w)] \right\}^{(j+i)\tau-1}, \end{aligned}$$

where $w = [\log(x) - \mu]/\sigma$ and $K = n!/[(i-1)!(n-i)!]$.

2.6 Inference

We consider the situation when the time-to-event is not completely observed and is subject to right censoring. Let C_i denote the censoring time. We observe $x_i = \min\{X_i, C_i\}$ and $\delta_i = I(X_i \leq C_i)$, where $\delta_i = 1$ if X_i is a time-to-event and $\delta_i = 0$ if it is right censored (for $i = 1, \dots, n$). Let c denote the parameter vector of the distribution of the time-to-event. Let X_i be a random variable following (2.6) with the vector of parameters $\gamma = (\mu, \sigma, \nu, \tau)^T$. From n pairs of times and censoring indicators $(x_1, \delta_1), \dots, (x_n, \delta_n)$, the log-likelihood function under non-informative censoring is given by

$$\begin{aligned} l(\gamma) &= r[\log(\tau\nu) - \log(\sigma\pi)] - \sum_{i \in F} \log(x_i) + \sum_{i \in F} \log \cosh(w_i) - \sum_{i \in F} \log [1 + \nu^2 \sinh^2(w_i)] \\ &+ (\tau-1) \sum_{i \in F} \log \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu \sinh(w_i)] \right\} + \sum_{i \in C} \log \left(1 - \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu \sinh(w_i)] \right\}^\tau \right), \end{aligned} \quad (2.24)$$

where r is the number of failures (uncensored observations).

We can obtain the MLE $\hat{\gamma}$ of γ by maximizing the log-likelihood (2.24) either directly in R using the `optim` function, in SAS using the `NLMixed` procedure and in other statistical software or by solving the nonlinear likelihood equations obtained by differentiating (2.24). The score functions for the parameters in γ are given by

$$\begin{aligned}
 U_{\mu}(\gamma) &= -\sum_{i \in F} \frac{\tanh(w_i)}{\sigma} + \sum_{i \in F} \frac{\nu^2 \sinh(2w_i)}{\sigma K_i} + (\tau - 1) \sum_{i \in F} \frac{\nu \cosh(w_i)}{\pi \sigma J_i K_i} + \sum_{i \in C} \frac{\tau \nu \cosh(w_i) J_i^{\tau-1}}{\pi \sigma K_i (J_i^{\tau} - 1)}, \\
 U_{\sigma}(\gamma) &= -\frac{r}{\sigma} - \sum_{i \in F} \frac{w_i}{\sigma} \tanh(w_i) + \sum_{i \in F} \frac{2\nu^2 w_i}{\sigma K_i} \sinh(w_i) \cosh(w_i) + (\tau - 1) \sum_{i \in F} \frac{\nu w_i}{\pi \sigma J_i K_i} \cosh(w_i) \\
 &\quad + \sum_{i \in C} \frac{\tau \nu w_i J_i^{\tau-1}}{\pi \sigma K_i (1 - J_i^{\tau})} \cosh(w_i), \\
 U_{\nu}(\gamma) &= \frac{r}{\nu} - \sum_{i \in F} \frac{2\nu \sinh^2(w_i)}{K_i} + (\tau - 1) \sum_{i \in F} \frac{\sinh(w_i)}{\pi J_i K_i} + \sum_{i \in C} \frac{\tau J_i^{\tau-1} \sinh(w_i)}{\pi K_i (J_i^{\tau} - 1)}
 \end{aligned}$$

and

$$U_{\tau}(\gamma) = \frac{r}{\tau} + \sum_{i \in F} \log(J_i) + \sum_{i \in C} \frac{J_i^{\tau}}{J_i^{\tau} - 1} \log(J_i),$$

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

The numerical maximization of the log-likelihood function (2.24) can also be performed in the GAMLSS package in R. The advantage of this package is that we can use many maximization methods, which will depend only on the current fitted model. When there are no explanatory variables or censored observations, we can use the `gamlssML` function for fitting (2.24) using a non-linear maximization algorithm. When we have 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 algorithms adopted in the presence of censored data are the RS and CG procedures. All methods and algorithms are described by Rigby and Stasinopoulos (2005) and Stasinopoulos and Rigby (2007) and they are available in the documentation of the GAMLSS package. The RS algorithm requires the first order derivatives of the logarithm of the density function (2.6) given in the above equations, and the second order derivatives. The RS method, different from the CG algorithm, does not use the cross derivatives, and thus it is faster for larger data sets. The second order derivatives can be determined numerically in the script discussed in Section 2.8.

Under standard regularity conditions, the asymptotic distribution of $(\hat{\gamma} - \gamma)$ is $N_4(\mathbf{0}, I(\gamma)^{-1})$, where $I(\gamma)$ is the expected information matrix. This asymptotic behavior holds if $I(\gamma)$ is replaced by $J(\hat{\gamma})$, i.e., the observed information matrix evaluated at the MLE $\hat{\gamma}$. Thus, the multivariate normal $N_4(\mathbf{0}, J(\hat{\gamma})^{-1})$ distribution can be used to construct approximate confidence intervals for the individual parameters.

Further, we can compute the maximum values of the log-likelihoods to obtain the likelihood ratio (LR) statistics for testing some sub-models of the ELSC distribution. For example, the test of $H_0 : \tau = 1$ versus $H : \tau \neq 1$ is equivalent to compare the LSC and ELSC distributions. In this case, the LR statistic is given by

$$w = 2\{l(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}) - l(\tilde{\mu}, \tilde{\sigma}, \tilde{\nu}, 0)\},$$

where $\hat{\mu}$, $\hat{\sigma}$, $\hat{\nu}$ and $\hat{\tau}$ are the MLEs under H and $\tilde{\mu}$, $\tilde{\sigma}$ and $\tilde{\nu}$ are the estimates under H_0 .

2.7 Simulation

We simulate the ELSC distribution (for $\mu = 4, \sigma = 0.1, \nu = 0.05, 0.6, 1.2$ and $\tau = 0.5, 1.5, 2$), considering bi-modality and unimodal forms, from equation (2.7) by using a random variable U having a uniform distribution in $(0, 1)$. We take $n=50, 150$ and 300 and, for each replication, we calculate the

MLEs $\hat{\mu}$, $\hat{\sigma}$, $\hat{\nu}$ and $\hat{\tau}$. We repeat this process 1,000 times and determine the average estimates (AEs), biases and means squared errors (MSEs). The results of the Monte Carlo study are given in Table 2.1. They indicate that the MSEs of the MLEs of μ , σ , ν and τ decay toward zero as the sample size increases, as expected under standard asymptotic theory.

Table 2.1. The AEs, biases and MSEs based on 1,000 simulations of the ELSC distribution for $\mu=4$ and $\sigma=0.1$, $\nu=0.05, 0.6, 1.2$ and $\tau=0.5, 1.5, 2$, and $n=50, 150$ and 300 .

n	Parameter	$\nu = 0.05$ and $\tau = 2$			$\nu = 0.6$ and $\tau = 2$			$\nu = 1.2$ and $\tau = 2$		
		AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
50	μ	4.001	0.001	0.001	2.913	-0.014	0.007	3.987	-0.013	0.003
	σ	0.097	-0.003	0.000	0.095	-0.005	0.001	0.099	-0.001	0.001
	ν	0.048	-0.002	0.001	0.635	0.035	1.371	1.321	0.121	0.433
	τ	2.050	0.050	0.143	2.913	0.913	42.345	2.884	0.884	7.379
150	μ	4.000	0.000	0.000	3.996	-0.004	0.003	3.989	-0.011	0.001
	σ	0.099	-0.001	0.000	0.098	-0.022	0.000	0.100	0.001	0.001
	ν	0.050	0.000	0.000	0.578	-0.022	0.026	1.209	0.009	0.093
	τ	2.014	0.014	0.045	2.181	0.181	1.051	2.368	0.368	1.044
300	μ	4.000	0.000	0.000	3.999	-0.001	0.002	3.996	-0.004	0.001
	σ	0.100	0.000	0.000	0.098	-0.002	0.000	0.100	0.001	0.001
	ν	0.050	0.000	0.000	0.580	-0.020	0.011	1.203	0.003	0.040
	τ	2.008	0.008	0.023	2.062	0.062	0.293	2.145	0.145	0.321
n	Parameter	$\nu = 0.05$ and $\tau = 1.5$			$\nu = 0.6$ and $\tau = 1.5$			$\nu = 1.2$ and $\tau = 1.5$		
		AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
50	μ	4.001	0.001	0.001	3.989	-0.011	0.006	3.990	-0.010	0.003
	σ	0.098	-0.002	0.001	0.097	-0.003	0.001	0.097	-0.003	0.001
	ν	0.050	0.001	0.001	0.581	-0.019	0.089	1.224	0.024	0.351
	τ	1.537	0.037	0.083	1.769	0.269	1.004	1.921	0.421	2.007
150	μ	4.001	0.001	0.001	3.995	-0.005	0.003	3.996	-0.004	0.001
	σ	0.099	-0.001	0.001	0.097	-0.003	0.001	0.101	0.001	0.001
	ν	0.050	0.001	0.001	0.578	-0.022	0.024	1.228	0.028	0.094
	τ	1.508	0.008	0.026	1.610	0.110	0.297	1.631	0.131	0.319
300	μ	4.000	0.001	0.001	3.998	-0.002	0.001	3.998	-0.002	0.001
	σ	0.100	0.001	0.001	0.099	-0.001	0.001	0.099	-0.001	0.001
	ν	0.050	0.001	0.001	0.583	-0.017	0.011	1.197	-0.003	0.040
	τ	1.508	0.008	0.013	1.550	0.050	0.129	1.562	0.062	0.107
n	Parameter	$\nu = 0.05$ and $\tau = 0.5$			$\nu = 0.6$ and $\tau = 0.5$			$\nu = 1.2$ and $\tau = 0.5$		
		AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
50	μ	3.998	-0.002	0.001	3.982	-0.018	0.008	4.003	0.003	0.003
	σ	0.097	-0.003	0.001	0.100	0.000	0.002	0.094	-0.006	0.002
	ν	0.049	-0.001	0.001	0.611	0.011	0.143	1.226	0.026	0.419
	τ	0.503	0.003	0.012	0.578	0.078	0.127	0.498	-0.002	0.075
150	μ	4.000	0.001	0.001	3.990	-0.010	0.003	4.006	0.006	0.001
	σ	0.099	-0.001	0.001	0.101	0.001	0.001	0.097	-0.003	0.001
	ν	0.049	-0.001	0.001	0.600	0.000	0.038	1.200	0.000	0.122
	τ	0.498	-0.002	0.004	0.538	0.038	0.040	0.485	-0.015	0.015
300	μ	4.000	0.001	0.001	3.996	-0.004	0.001	4.002	0.002	0.001
	σ	0.100	0.001	0.001	0.101	0.001	0.001	0.099	-0.001	0.001
	ν	0.050	0.001	0.001	0.602	0.002	0.018	1.205	0.005	0.054
	τ	0.500	0.001	0.002	0.516	0.016	0.015	0.493	-0.007	0.007

We conclude from the figures in Table 2.1 that the AEs of the parameters tend to be closer to the true parameters 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 be oftentimes improved by using bias adjustments to these estimators. Approximations to the their biases in simple models may be determined analytically. Bias correction typically does a very good job for correcting the MLEs. However, it may also increase the MSEs. Whether bias correction is useful in practice depends basically on the shape of the bias function and on the variance of the MLE. In order to improve the accuracy of these estimators using analytical bias reduction one needs to obtain several cumulants of log-likelihood derivatives, which are notoriously cumbersome for the proposed model. We illustrate the convergence in Figures 2.6 and 2.7, where the true densities are given at selected parameter values and the density functions are computed at the AEs given in Table 2.1 for some sample sizes and $\nu = 0.05$ and $\nu = 0.6$, respectively. In Figures 2.8 and 2.9, we present the estimated densities based on 1,000 samples of the AEs of the parameters μ , σ , τ for $\nu = 0.05$ and $\nu = 0.6$, respectively, and $n = 50, 150$ and 300 . These plots are in agreement with the standard asymptotic theory for the MLEs.

Figure 2.6. Some ELSC density functions at the true parameter values and at the AEs for $\mu=4$, $\sigma=0.1$, $\nu=0.05$ and $\tau=2$ when: (a) $n=50$; (b) $n=150$; (c) $n=300$.

Figure 2.7. Some ELSC density functions at the true parameter values and at the AEs for $\mu=4$, $\sigma=0.1$, $\nu=0.6$ and $\tau=2$ when: (a) $n=50$; (b) $n=150$; (c) $n=300$.

Figure 2.8. Estimated densities from 1,000 samples for $n = 50, 150, 300$ of the parameters: (a) $\mu = 4$; (b) $\sigma = 0.1$; (c) $\nu = 0.05$; (d) $\tau = 2$ (based on selected parameter values in Table 1 for $\nu = 0.05$).

Figure 2.9. Estimated densities from 1,000 samples for $n = 50, 150, 300$ of the parameters: (a) $\mu = 4$; (b) $\sigma = 0.1$; (c) $\nu = 0.6$; (d) $\tau = 2$ (based on selected parameter values in Table 1 for $\nu = 0.6$).

2.8 Applications

In this section, we provide three applications to real data to prove empirically the flexibility of the ELSC and LSC models. The computations are performed using the `gamlss` subroutine in the R software. In the first application, we give an application for bimodal data comparing the ELSC and LSC models with other models implemented in `gamlss`. In the second application, we show the flexibility of the distribution for censored data and, in the third application, we study the adequacy of the LSC model.

Recently, Cordeiro *et al.* (2014) proposed the McDonald-Weibull (McW) model with scale parameter $\lambda > 0$, shape parameter $\gamma > 0$ and three extra shape parameters $a > 0$, $b > 0$ and $c > 0$. We focus on this model since it extends various distributions previously discussed in the lifetime literature, such as the beta Weibull (BW) (Lee *et al.*, 2007) (for $c = 1$), Kumaraswamy Weibull (KwW) (Cordeiro *et al.*, 2010) (for $a = c$), exponentiated Weibull (EW) (Mudholkar *et al.*, 1995) (for $b = c = 1$), Weibull (for $a = b = c = 1$) and other distributions. Besides of its flexibility, the McW model can take bimodal forms and thus is a competitive model for the ELSC distribution.

All computations in this section are performed using the `gamlss` subroutine in R and the scripts are described in Section 2.9.

2.8.1 Eruption data

First, we provide an analysis of some data on the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. The data consist of $n = 299$ pairs of measurements referring to the times between the starts of successive eruptions. These data were collected continuously from August 1st until August 15th, 1985; see Azzalini and Bowman (1990) for more details.

We compute the Hartigans' Dip statistic D and its p -value for the test for unimodality. For i.i.d. random variables, the null hypothesis is that X_i has a unimodal distribution. Consequently, the alternative hypothesis is non-unimodal, i.e., at least bimodal. The Dip test can be obtained using a function `dip.test` available in "diptest" R package. More details about the dip test can be obtained in Hartigan and Hartigan (1985). Applying the Dip test to verify that a unimodal distribution would be appropriate to fit the eruption data gives $D = 0.039$ with the p -value 0.002. So, we reject the null hypothesis in favor of a bimodal distribution.

Further, we compare the fits of the ELSC and LSC models with the models available in the `gamlss.family` package. The `fitDist(..., type=c("realplus"))` function is used to fit all relevant parametric distributions. The Box-Cox power exponential (BCPEo) distribution is selected as the best model. For details on the distributions available in the package, see Stasinopoulos *et al.* (2014). Table 2.2 lists the MLEs (and the corresponding standard errors in parentheses) of the model parameters and the values of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) statistics for the fitted models.

We also evaluate the Cramér-von Mises (W^*) and Anderson-Darling (A^*) statistics described by Chen and Balakrishnan (1995). From a random sample $x_1, \dots, x_n$ with empirical distribution function $F_n(x)$, the main objective is to test if the sample comes from a specific distribution. The W^* and A^* statistics are given by

$$\begin{aligned} W^* &= \left(n \int_{-\infty}^{+\infty} \{F_n(x) - F(x; \hat{\gamma}_n)\}^2 dF(x; \hat{\gamma}_n) \right) \left(1 + \frac{0.5}{n} \right) = W^2 \left(1 + \frac{0.5}{n} \right), \\ A^* &= \left(n \int_{-\infty}^{+\infty} \frac{\{F_n(x) - F(x; \hat{\gamma}_n)\}^2}{\{F(x; \hat{\gamma}_n)(1 - F(x; \hat{\gamma}_n))\}} dF(x; \hat{\gamma}_n) \right) \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2} \right), \\ &= A^2 \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2} \right), \end{aligned}$$

respectively, where $F_n(x)$ is the empirical distribution function and $F(x; \hat{\gamma}_n)$ is the postulated distribution function evaluated at the MLE $\hat{\gamma}_n$ of γ . The W^* and A^* statistics measure the differences of $F_n(x)$ and $F(x; \hat{\gamma}_n)$. Thus, the lower their values, the more evidence that $F(x; \hat{\gamma}_n)$ generates the sample.

The figures in Table 2 indicate that the ELSC model has the lowest AIC and BIC values among those values of the fitted models, and therefore it could be chosen as the best model. Further, the SEs of the estimates for all fitted models are quite small.

Table 2.2. MLEs of the model parameters for the eruption data, the corresponding SEs and the AIC and BIC statistics.

Model	μ	σ	ν	τ	AIC	BIC	W^*	A^*
ELSC	4.153 (0.008)	0.069 (0.056)	0.089 (0.193)	1.728 (0.078)	2328.23	2343.03	0.08	0.70
LSC	4.193 (0.007)	0.065 (0.057)	0.101 (0.201)	- -	2368.26	2379.36	0.32	2.18
BCPEo	70.675 (0.014)	0.191 (0.032)	0.966 (0.271)	4.973 (0.143)	2387.22	2402.02	0.82	4.36

Formal tests for the extra skewness parameters in the ELSC model can be based on the LR statistic described in Section 2.6. Applying the LR statistic to the eruption data, we reject the null hypothesis $H_0 : \tau = 1$ in favor of the ELSC distribution. The value of the LR statistic is $w = 42.032$ with the p -value < 0.001 .

More information is provided by a visual comparison of the histogram of the data with the fitted density functions. The plots of the fitted ELSC, LSC and BCPEo densities and their cdfs are displayed in Figure 2.10. The plot of the ELSC hazard rate in Figure 2.11 reveals that this function has a bimodal shape, small at the first mode and large at the second mode.

Figure 2.10. Estimated (a) densities and (b) cdfs for the ELSC, LSC and BCPEo models fitted to the eruption data.

2.8.2 Efron data

Second, we consider the data from a two-arm clinical trial discussed earlier by Efron (1988). Efron noted that the empirical hazard functions for both samples start near zero, suggesting an initial high-risk period at the beginning, a decline for a while, and then stabilization after about one year. He developed and illustrated a methodology for analyzing the data using a combination of techniques of quantal response analysis and the spline regression methods. Specifically, Efron's data from a head

Figure 2.11. Estimated hrf for the ELSC distribution for eruption data.

and neck cancer clinical trial consist of survival times of 51 patients in arm A who were given radiation therapy and 45 patients in arm B who were given radiation plus chemotherapy. Nine patients in arm A and 14 patients in arm B were lost to follow-up and were regarded as censored.

Cordeiro *et al.* (2014) fitted the McW regression model to these data and noted that it provides a good fit. Here, we consider only the survival times in days x_i and compare the results of the fits of the McW, ELSC and LSC models. Table 2.3 gives the MLEs (and the corresponding standard errors in parentheses) of the parameters and the values of the AIC and BIC statistics. They indicate that the ELSC model has the lowest values of these statistics among the values of the other fitted models, and therefore it could be chosen as the best model.

Table 2.3. MLEs of the model parameters for Efron data, the corresponding SEs (given in parentheses) and the AIC and BIC statistics.

Model	μ	σ	ν	τ		AIC	BIC
ELSC	4.788 (0.083)	2.080 (0.135)	2.794 (0.129)	2.308 (0.097)		1063.9	1074.1
LSC	6.141 (0.102)	0.494 (0.061)	0.215 (0.151)	1 -		1074.4	1082.1
	λ	γ	a	b	c	AIC	BIC
McW	0.092 (0.028)	0.101 (0.008)	74.352 (0.655)	21.126 (0.192)	0.067 (0.001)	1088.5	1101.3
BW	0.281 (0.106)	0.062 (0.005)	167.450 (0.406)	60.159 (0.177)	1 -	1086.1	1096.3

By comparing the fits of the ELSC and LSC models using the LR statistic, we reject the null hypothesis $H_0 : \tau = 1$ in favor of the ELSC distribution. The LR statistic is $w = 12.552$ with the p -value < 0.001 . Next, we compare the fits of the McW and BW models using the LR statistic. Applying the LR statistic for testing the null hypothesis $H_0 : c = 1$, we obtain $w = 0.00039$ with the p -value almost one. So, we could not reject the BW distribution to fit these data.

The plots of the fitted ELSC, LSC and BW densities and their estimated survival functions are displayed in Figure 2.12 for the current data ignoring censored observations. Clearly, the ELSC density provides a closer fit to the histogram of the data and the corresponding estimated survival function to the empirical survival function than the other models. The plot of the ELSC hrf in Figure 2.13 reveals that it has a modal shape.

Figure 2.12. (a) Estimated ELSC, LSC and BW densities for Efron data. (b) Estimated ELSC and LSC survival functions and the empirical survival for Efron data.

Figure 2.13. Estimated ELSC hazard function for Efron data.

2.8.3 Entomology data

Third, we consider the data from a study carried out at the Department of Entomology of the Luiz de Queiroz School of Agriculture, University of São Paulo, which aim to assess the longevity of the mediterranean fruit fly (*ceratitis capitata*). The need for this fly to seek food just after emerging from the larval stage has permitted the use of toxic baits for its management in Brazilian orchards for at least fifty years. This pest control technique consists of using small portions of food laced with an insecticide, generally an organophosphate, that quickly kills the flies, instead of using an insecticide alone. Recently, there have been reports of the insecticidal effect of extracts of the neem tree leading to proposals to adopt various extracts (aqueous extract of the seeds, methanol extract of the leaves and dichloromethane extract of the branches) to control pests such as the mediterranean fruit fly. For more details, see Silva *et al.* (2013).

The response variable in the experiment is the lifetime of the adult flies in days after exposure to the treatments. The experimental period was set at 51 days, so that the numbers of larvae that survived beyond this period are considered as censored observations. The total sample size is $n = 72$ because four cases are lost. Therefore, the variables used in this study are: x_i -lifetime of *ceratitis capitata* adults in days and δ_i -censoring indicator.

Recently, Lanjoni (2013) fitted the Burr XII geometric type II (BXIIGII) distribution to these data and noted that it gives a better fit than the special Burr XII model. Now, we compare the McW

and BXIIGII distributions and some of their sub-models with the ELSC and LSC models. For some fitted models, Table 2.4 provides the MLEs (and the corresponding standard errors in parentheses) of the parameters and the values of the AIC and BIC statistics. The computations are performed using the `gamlss` subroutine in R. They indicate that the LSC model has the lowest AIC and BIC values among those values of the fitted models, and therefore it could be chosen as the best model. The LSC model is not able to capture asymmetry but it has the bi-modality characteristic.

Table 2.4. MLEs of the model parameters for the entomology data, the corresponding SEs (given in parentheses) and the AIC and BIC statistics.

Model	μ	σ	ν	τ	AIC	BIC	
ELSC	3.018 (0.027)	0.852 (0.091)	3.367 (0.107)	0.907 (0.075)	1249.0	1261.5	
LSC	2.998 (0.029)	0.946 (0.101)	3.592 (0.106)	1 -	1247.7	1257.1	
	s	c	k	p	AIC	BIC	
BXIIGII	14.353 (8.175)	1.164 (0.389)	4.414 (2.532)	0.981 (0.0211)	1270.1	1282.7	
BXII	34.423 (10.386)	2.214 (0.232)	2.676 (1.284)	1 -	1282.7	1292.1	
	λ	γ	a	b	c	AIC	BIC
McW	0.079 (0.007)	1.718 (0.223)	0.883 (0.313)	0.329 (0.114)	0.049 (0.013)	1290.0	1305.8
BW	0.055 (0.017)	1.608 (0.226)	1.240 (0.314)	0.688 (0.313)	1 -	1289.7	1302.3
KwW	0.015 (0.004)	1.133 (0.447)	1 -	8.787 (0.299)	1.776 (0.920)	1288.9 -	1301.5
EW	0.044 (0.007)	1.587 (0.275)	1.254 (0.368)	1 -	1 -	1287.5	1296.9
Weibull	0.0400 (0.002)	1.797 (0.111)	1 -	1 -	1 -	1286.1	1292.4

In order to assess if the model is appropriate, Figure 2.14a displays the empirical and estimated cumulative distributions for the fitted ELSC and LSC models to the current data. Further, Figure 2.14b gives the plots of the empirical survival function and the estimated ELSC and LSC survival functions. They indicate the LSC model provides a good fit to these data. Further, using the LR statistic to compare the fits of these models, i.e. for testing the null hypothesis $H_0 : \tau = 1$, we obtain $w = 0.748$ with the p -value= 0.387 and then we could accept the LSC distribution. The plot of its hrf in Figure 2.15 reveals a modal shape.

2.9 Program description

The ELSC model is implemented in the `gamlss` function, which is fully documented in the `gamlss` package (Stasinopoulos and Rigby, 2007). Here, we will omit several functions for the `gamlss` package and present only the functions related to the ELSC distribution and its fit to a data set. The computational codes for the ELSC model can be downloaded from <http://goo.gl/yzvoIZ>. The cdf (2.5) and pdf (2.6) can be obtained using `dELSC` and `pELSC` functions, respectively. The qf given by (2.7) can be obtained using the `qELSC` function and samples of the ELSC model can be generated using the `rELSC` function. We can use the functions listed above for the LSC sub-model by setting $\tau = 1$ with the `tau.fix=TRUE` function. To optimize the computational time, we can change the initial values of the parameters using the `parameter.fix` function. Otherwise, we can increase the number of interactions using the `n.cyc` function. The fit of the ELSC model to censored data can be performed using the

Figure 2.14. (a) Estimated ELSC and LSC cdfs for entomology data. (b) Estimated ELSC and LSC survival functions and the empirical survival for the entomology data.

Figure 2.15. Estimated LSC hazard function for entomology data.

additional package `gamlss.cens`. The structure of the `gamlss` function is familiar to users of the R syntax (the `glm` function, in particular).

2.10 Conclusions

The paper proposes the exponentiated log-sinh Cauchy (ELSC) distribution that can be used as an alternative to mixture distributions in modeling bimodal data. Various mathematical properties of the ELSC distribution are investigated. We show that it can accommodate various shapes of the skewness, kurtosis and bi-modality. Its model parameters are estimated by maximum likelihood. Some numerical experiments reveal that the maximum likelihood estimation procedure performs well. Three real data examples prove empirically that the ELSC distribution is very flexible, parsimonious, and a competitive model that deserves to be added to existing distributions in modeling bimodal data. The ELSC model can be fitted using the `gamlss` package described to facilitate its practical use by researchers from other areas.

References

Azzalini, A. and Bowman, A.W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics*, 357–365.

- Chen, G. and Balakrishnan, N. (1995). A general purpose approximate goodness-of-fit test. *Journal of Quality Technology*, **27**, 154–161.
- Comtet, L. (1974). *Advanced Combinatorics*. D. Reidel Publishing Co., Dordrecht.
- Cooray, K. (2013). Exponentiated Sinh Cauchy Distribution with Applications. *Communications in Statistics-Theory and Methods*, **42**, 3838–3852.
- Cordeiro, G.M., Ortega, E.M.M. and Nadarajah, S. (2010). The Kumaraswamy Weibull distribution with application to failure data. *Journal of the Franklin Institute*, **347**, 1399–1429.
- Cordeiro, G.M., Hashimoto, E.M. and Ortega, E.M. (2014). The McDonald Weibull model. *Statistics*, **48**, 256–278.
- de Castro, M., Cancho, V.G. and Rodrigues, J. (2010). A hands-on approach for fitting long-term survival models under the GAMLSS framework. *Computer methods and programs in biomedicine*, **97**, 168–177.
- Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, **83**, 414–425.
- Galton, F. (1883). *Inquiries Into the Human Faculty & Its Development*.
- Gupta, R.D. and Kundu, D. (2001). Exponentiated exponential family: an alternative to gamma and Weibull distributions. *Biometrical journal*, **43**, 117–130.
- Hartigan, J.A. and Hartigan, P.M. (1985). The dip test of unimodality. *The Annals of Statistics*, 70–84.
- Lanjoni, B.R. (2013). O modelo Burr XII geométrico: propriedades e aplicações. Master's Dissertation, Escola Superior de Agricultura Luiz de Queiroz, University of São Paulo, Piracicaba. Retrieved 2015-05-27, from <http://www.teses.usp.br/teses/disponiveis/11/11134/tde-17122013-085812/>.
- Lee, C., Famoye, F. and Olumolade, O. (2007). Beta Weibull distribution: some properties and applications to censored data. *Journal of Modern Applied Statistical Methods*, **6**, 173–186.
- Moors, J.J.A. (1988). A quantile alternative for kurtosis. *The statistician*, 25–32.
- Mudholkar, G.S., Srivastava, D.K. and Freimer, M. (1995). The exponentiated Weibull family: A reanalysis of the bus-motor-failure data. *Technometrics*, **37**, 436–445.
- 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.
- Rodrigues, J., de Castro, M., Cancho, V.G. and Balakrishnan, N. (2009). COM–Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, **139**, 3605–3611.
- Silva, M.A., Bezerra-Silva, G.C.D., Vendramim, J.D. and Mastrangelo, T. (2013). Sublethal effect of neem extract on Mediterranean fruit fly adults. *Revista Brasileira de Fruticultura*, **35**, 93–101.
- 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.
- Stasinopoulos, D.M., Rigby, R.A., Akantziliotou, C., Heller, G., Ospina, R., Voudouris, M.D., ... and Stasinopoulos, M.M. (2014). Package “gamlss.dist”.
- Team, R.C. (2000). *R Language Definition*.

3 NEW REGRESSION MODEL WITH FOUR REGRESSION STRUCTURES AND COMPUTATIONAL ASPECTS

Abstract: A new general class of exponentiated sinh Cauchy regression models for location, scale and shape parameters is introduced and studied. It may be applied to censored data and used more effectively in survival analysis when compared with the usual models. For censored data, we employ a frequentist analysis for the parameters of the proposed model. Further, for different parameter settings, sample sizes and censoring percentages, various simulations are performed. The extended regression model is very useful for the analysis of real data and could give more adequate fits than other special regression models.

Keywords: Exponentiated sinh Cauchy regression model; diagnostics analysis; GAMLSS; survival analysis.

3.1 Introduction

The Weibull, log-normal, log-logistic and Birnbaum-Saunders regression models are usually applied in science and engineering to model lifetime data for which linear functions of unknown parameters are adapted to explain the phenomena under study. However, it is well-known that several phenomena are not always in agreement with the usual model due to lack of asymmetry, bimodality or the presence of heavily and lightly tailed distributions. In order to deal with this problem, some proposals have been made in literature with more flexible classes of distributions. We work with the exponentiated sinh Cauchy distribution because of its great flexibility to fit asymmetric and bimodal data.

A large number of new distributions to extend well-known distributions and to provide flexibility in modeling data has been investigated in the last years. In this context, Gupta *et al.* (1998) pioneered a generalization of the standard exponential distribution called the exponentiated exponential (Exp-E) distribution. The exponentiated class of distributions (Gupta and Kundu, 2001) has cumulative distribution function (cdf) given by

$$F(t) = G(t)^\tau, \quad (3.1)$$

where $G(t)$ represents the baseline cdf and $\alpha > 0$ denotes the shape parameter. By differentiating (3.1), the corresponding probability density function (pdf) becomes

$$f(x) = \tau G(t)^{\tau-1} g(t), \quad (3.2)$$

where $g(t)$ denotes the baseline pdf.

For modeling a lifetime $T > 0$, Ramires *et al.* (2016) used the *log-sinh Cauchy* (LSC) distribution for the baseline in (3.2) by defining the four-parameter *exponentiated log-sinh Cauchy* (ELSC) distribution, whose pdf (for $t > 0$) is given by

$$f(t; \mu, \sigma, \nu, \tau) = \frac{\tau \nu}{t \sigma \pi} \frac{\cosh\left(\frac{\log(t) - \mu}{\sigma}\right)}{\left[\nu^2 \sinh^2\left(\frac{\log(t) - \mu}{\sigma}\right) + 1\right]} \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh\left(\frac{\log(t) - \mu}{\sigma}\right) \right] \right\}^{\tau-1}, \quad (3.3)$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are the location and scale parameters, respectively, $\nu > 0$ is the symmetry parameter, which characterizes the bimodality of the distribution, and $\tau > 0$ is the skewness parameter. The distribution of the logarithm $Y = \log(T)$ is called the *exponentiated sinh Cauchy* (ESC) distribution, whose cdf (for $y \in \mathbb{R}$) is given by

$$F(y; \mu, \sigma, \nu, \tau) = \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh\left(\frac{y - \mu}{\sigma}\right) \right] \right\}^\tau. \quad (3.4)$$

The pdf and survival function corresponding to (3.4) are given by

$$f(y; \mu, \sigma, \nu, \tau) = \frac{\tau \nu}{\sigma \pi} \frac{\cosh\left(\frac{y-\mu}{\sigma}\right)}{[\nu^2 \sinh^2\left(\frac{y-\mu}{\sigma}\right) + 1]} \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh\left(\frac{y-\mu}{\sigma}\right) \right] \right\}^{\tau-1} \quad (3.5)$$

and

$$S(y; \mu, \sigma, \nu, \tau) = \frac{(2\pi)^\tau - \left\{ \pi + 2 \arctan \left[\nu \sinh\left(\frac{y-\mu}{\sigma}\right) \right] \right\}^\tau}{(2\pi)^\tau}, \quad (3.6)$$

respectively. The ESC distribution (3.5) was first introduced by Cooray (2013) to modeling symmetric, right and left skewed and bimodal data sets. For $\tau = 1$, the sinh Cauchy (SC) distribution is just a special case of (3.5).

In this paper, we propose a general class of regression models, where the mean, dispersion, asymmetry and bimodal parameters vary across observations through regression structures, assuming that the model errors follow the ESC distribution, which may be a useful alternative for modeling the four existing types of failure rate functions. The inferential component is carried out using the asymptotic distribution of the maximum likelihood estimators (MLEs). We also present methodologies to detect influential subjects with censored data and residual analysis for the proposed model. The script used to fit the ESC model, which is implemented in the R software environment (R Core Team, 2015), is given in the Section 3.9.

The sections are organized as follows. In Section 3.2, we derive a power series for the quantile function (qf) and give explicit expressions for the moments. We propose an ESC regression model for modeling simultaneously the location, scale, bimodality and asymmetry parameters for censored data and discuss inferential issues in Section 3.3. Section 3.4 contains some Monte Carlo simultaneously on the finite sample behavior of the MLEs. In Section 3.5, we assess the behavior of the MLEs of the parameters in the ESC regression model when it is poorly specified. In Section 3.6, we discuss some diagnostic measures for three perturbation schemes, case-deletion and generalized leverage method. The residuals from a fitted model using the martingale residual and martingale-type residual are also presented in this section. Applications to two real data sets are addressed in Section 3.7 to illustrate the flexibility of the proposed class of regression models for censored and uncensored data. Finally, Section 3.8 offers some conclusions.

3.2 Properties of the standardized ESC distribution

In this section, we study some properties of the standard ESC random variable defined by $Z = (Y - \mu)/\sigma$. The density function of Z (for $z \in R$) reduces to

$$f(z; \nu, \tau) = \tau g_{SC}(z) G_{SC}(z)^{\tau-1} = \frac{\tau \nu}{\pi} \frac{\cosh(z)}{\nu^2 \sinh^2(z) + 1} \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan [\nu \sinh(z)] \right\}^{\tau-1}, \quad (3.7)$$

where $G_{SC}(z)$ and $g_{SC}(z)$ denote the cdf and pdf of standard SC distribution given by

$$G_{SC}(z) = \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan [\nu \sinh(z)] \right\} \quad \text{and} \quad g_{SC}(z) = \frac{\nu}{\pi} \frac{\cosh(z)}{\nu^2 \sinh^2(z) + 1}, \quad (3.8)$$

respectively.

Plots of the density function (3.7) for selected parameter values are displayed in Figure 3.1. Equation (3.7) for the standardized ESC distribution will be used in Section 3.3.1 to specify the error distribution of the proposed regression model.

Figure 3.1. Plots of the density function (3.7) for some values of τ : (a) $\nu = 0.3$; (b) $\nu = 0.8$.

3.2.1 Expansion of the quantile function

Inverting $F(y) = u$ in (3.4) gives the qf of Y

$$Y = Q_Y(u) = \mu + \sigma \operatorname{arcsinh} \left\{ \frac{1}{\nu} \tan \left[\pi \left(u^{1/\tau} - 0.5 \right) \right] \right\}. \quad (3.9)$$

The qf $Q_Z(u)$ of Z , which has the standardized ESC density function (3.7), can be obtained from (3.9) with $\mu = 0$ and $\sigma = 1$. The qf of the standardized SC distribution, say $Q_{SC}(u)$, also follows (3.9) with $\mu = 0$ and $\sigma = \tau = 1$ and it will be used to demonstrate some properties of Z in the following sections.

We can use (3.9) for simulating ESC or standardized ESC random variables by setting u as a uniform random variable in the interval $(0, 1)$. The qf is widely used to determine some mathematical properties like moments, generating function, Galton's skewness and Moors's kurtosis. Recently, Ortega *et al.* (2016) used the qf to demonstrate some properties of the log-odds Birnbaum-Saunders model and Cordeiro *et al.* (2016) presented those for the generalized odd half-Cauchy family.

Next, we derive a power series for the qf of Z . Expanding (3.9) in Mathematica in a power series, considering $\mu = 0$ and $\sigma = 1$, we have

$$Q_Z(u) = \sum_{k=0}^{\infty} c_k (u^{1/\tau} - 0.5)^{2k+1},$$

where $c_k = \frac{b_k}{(2k+1)!} \left(\frac{\pi}{\nu} \right)^{2k+1}$ and $b_0 = 1, b_1 = 2\nu^2 - 1, b_2 = 16\nu^4 - 20\nu^2 + 9, b_3 = 272\nu^6 - 616\nu^4 + 630\nu^2 - 225, b_4 = 7936\nu^8 - 28160\nu^6 + 48384\nu^4 - 37800\nu^2 + 11025, \dots$

By expanding the binomial term, the last equation reduces to

$$Q_Z(u) = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \frac{(-1)^{2k+1-j} u^{j/\tau}}{2^{2k+1-j}} \binom{2k+1}{j} c_k.$$

Finally, changing $\sum_{k=0}^{\infty} \sum_{j=0}^{\infty}$ by $\sum_{j=0}^{\infty} \sum_{k=j}^{\infty}$, we obtain

$$Q_Z(u) = \sum_{j=0}^{\infty} p_j u^{j/\tau}, \quad (3.10)$$

where the coefficients

$$p_j = \sum_{k=j}^{\infty} (-0.5)^{2k+1-j} \binom{2k+1}{j} c_k \quad (3.11)$$

can be determined using e.g. Mathematica, Maple, R and Sage.

3.2.2 Moments

Let $\mu'_s = E(Z^s)$ be the s th ordinary moment of Z with pdf (3.7). We have

$$\mu'_s = \tau \int_{-\infty}^{\infty} z^s g_{SC}(z) G_{SC}(z)^{\tau-1} dz = \tau \int_0^1 Q_{SC}(u)^s u^{\tau-1} du.$$

Replacing $Q_{SC}(u)$ (eq. (3.10) when $\tau = 1$) in the last equation, we obtain

$$\mu'_n = \tau \int_0^1 \left(\sum_{j=0}^{\infty} p_j u^j \right)^s u^{\tau-1} du. \quad (3.12)$$

Henceforth, we use an equation by Gradshteyn and Ryzhik (2007) for a power series raised to a positive integer n

$$\left(\sum_{i=0}^{\infty} a_i u^i \right)^n = \sum_{i=0}^{\infty} b_{n,i} u^i, \quad (3.13)$$

where the coefficients $b_{n,i}$ (for $i = 1, 2, \dots$) are easily determined from the recurrence equation

$$b_{n,i} = (i a_0)^{-1} \sum_{m=1}^i [m(n+1) - i] a_m b_{n,i-m},$$

and $b_{n,0} = a_0^n$. The coefficient $b_{n,i}$ can be determined numerically from the quantities $a_0, \dots, a_i$.

Based on equation (3.13), equation (3.12) can be rewritten as

$$\mu'_n = \tau \sum_{j=0}^{\infty} e_{s,j} \int_0^1 u^{j+\tau-1} du = \sum_{j=0}^{\infty} \frac{\tau}{\tau+j} e_{s,j}, \quad (3.14)$$

where $e_{s,j} = \frac{1}{j! p_0} \sum_{m=1}^j [m(s+1) - j] p_m e_{s,j-m}$, $e_{s,0} = p_0^s$, and p_0 and p_m are obtained by (3.11).

The skewness and kurtosis measures can be calculated from the ordinary moments using well-known relationships. Plots of the skewness and kurtosis of Z are displayed in Figures 3.2 and 3.3 for selected values of τ as functions of ν and for selected values of ν as functions of τ , respectively.

Figure 3.2. Skewness of the ESC distribution: (a) Function of ν for some values of τ . (b) Function of τ for some values of ν .

3.3 The ESC regression model

In many practical applications, the lifetimes are affected by explanatory variables such as blood pressure, weight, cholesterol level and many others. Parametric models for estimating univariate survival

Figure 3.3. Kurtosis of the ESC distribution: (a) Function of ν for some values of τ . (b) Function of τ for some values of ν .

functions and for the censored data regression problems are widely used. When the parametric models provide good fits to lifetime data, they tend to provide more precise estimates for the quantities of interest because these estimates are based on fewer parameters. Recently, several regression models have been proposed in literature by considering the class of location models. For example, Hashimoto *et al.* (2012) proposed the log-Burr XII regression model for grouped survival data, Ortega *et al.* (2013) presented the log-beta Weibull regression model for predicting recurrence of prostate cancer, Ortega *et al.* (2015) studied a power series beta Weibull regression model for predicting breast carcinoma, etc.

A disadvantage of the class of location model is that the variance, skewness, bimodality, kurtosis and other parameters can not be modelled explicitly in terms of explanatory variables but implicitly through their dependence on the location parameter. As an alternative, the generalized additive models for location, scale and shape (GAMLSS) (Rigby and Stasinopoulos, 2005), where the systematic part of the model is expanded to allow not only the location but all the parameters of the conditional distribution of Y to be modelled as parametric functions of explanatory variables, become widely used. In this sense, we introduce the ESC regression model following the GAMLSS set-up.

3.3.1 Definition

Let $\theta^T = (\mu, \sigma, \nu, \tau)$ denote the vector of parameters of the pdf (3.5). We consider that independent observations y_i conditional on θ_i (for $i = 1, \dots, n$), with pdf $f(y_i; \theta_i)$, where $\theta_i^T = (\mu_i, \sigma_i, \nu_i, \tau_i)$ is a parameter vector related to the response variable.

Based on the ELSC distribution, we propose a linear regression model linking the response variable y_i and the explanatory variable by

$$y_i = \mu_i + \sigma_i z_i, \quad i = 1, \dots, n, \quad (3.15)$$

where the random error z_i follows the density function $f(z_i; \nu_i, \tau_i)$ given by (3.7) and $Z_i = (Y_i - \mu_i)/\sigma_i$. We define the parameter vector θ using appropriate link functions as

$$\theta = \begin{bmatrix} \mu \\ \sigma \\ \nu \\ \tau \end{bmatrix} = \begin{bmatrix} g_1(\mathbf{X}_1\beta_1) \\ g_2(\mathbf{X}_2\beta_2) \\ g_3(\mathbf{X}_3\beta_3) \\ g_4(\mathbf{X}_4\beta_4) \end{bmatrix} \text{ or } \theta_i = \begin{bmatrix} \mu_i \\ \sigma_i \\ \nu_i \\ \tau_i \end{bmatrix} = \begin{bmatrix} g_1(\beta_{01} + x_1[i, 2]\beta_{11} + \dots + x_1[i, p_1 + 1]\beta_{p_11}) \\ g_2(\beta_{02} + x_2[i, 2]\beta_{12} + \dots + x_2[i, p_2 + 1]\beta_{p_22}) \\ g_3(\beta_{03} + x_3[i, 2]\beta_{13} + \dots + x_3[i, p_3 + 1]\beta_{p_33}) \\ g_4(\beta_{04} + x_4[i, 2]\beta_{14} + \dots + x_4[i, p_4 + 1]\beta_{p_44}) \end{bmatrix}, \quad (3.16)$$

where p_k denotes the number of explanatory variables related to the k th parameter, $g_1(\cdot)$ is an injective and twice continuously differentiable function, $g_k(\cdot)$ (for $k = 2, 3, 4$) is a known positive continuously differentiable function containing values of the explanatory variables, $\beta_k = (\beta_{0k}, \beta_{1k}, \dots, \beta_{p_k k})^T$ is a parameter vector of length $(p_k + 1)$, $\mathbf{X}_k$ is a known model matrix of order $n \times (p_k + 1)$ and $x_k[i, p_k]$

are the elements of the matrix $\mathbf{X}_k$. The total number of parameters to be estimated is given by $p = p_1 + p_2 + p_3 + p_4 + 4$. Note that we assume that four parameters μ_i , σ_i , ν_i and τ_i vary across observations through regression structures. For the following sections, we shall consider the identity link function for $g_1(\cdot)$ and the logarithmic link function for $g_k(\cdot)$ (for $k = 2, 3, 4$).

The sinh Cauchy (SC) regression model is obtained as a special case of (3.15) when $\tau_i = 1$. The class of location is obtained when $p_2 = p_3 = p_4 = 0$. For $p_3 = p_4 = 0$, $p_1 \neq 0$ and $p_2 \neq 0$, we also obtain the regression model with heteroscedastic errors, which can be used as an alternative to transformation of the response variable. However, the choice of parameters to be modeled by explanatory variables will depend on the data set.

3.3.2 Estimation

Consider a sample of n -independent observations, where each random response is defined by $y_i = \min[\log(t_i), \log(c_i)]$. We assume non-informative censoring and that the observed lifetimes and censoring times are independent. Let F and C be the sets of individuals for which y_i is the log-lifetime or log-censoring, respectively. The total log-likelihood function for the model parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})^T$ from model (3.15) is given by $l(\boldsymbol{\theta}) = \sum_{i \in F} \log f(y_i; \boldsymbol{\theta}_i) + \sum_{i \in C} \log S(y_i; \boldsymbol{\theta}_i)$, where $f(y_i; \boldsymbol{\theta}_i)$ is the density function in (3.5) and $S(y_i; \boldsymbol{\theta}_i)$ is the survival function in (3.6). The log-likelihood function for $\boldsymbol{\theta}$ reduces to

$$\begin{aligned} l(\boldsymbol{\theta}) = & - \sum_{i \in F} \log [1 + \nu_i^2 \sinh^2(z_i)] + \sum_{i \in F} \log \cosh(z_i) + \sum_{i \in F} (\tau_i - 1) \log \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu_i \sinh(z_i)] \right\} \\ & + \sum_{i \in F} \log(\tau_i \nu_i) - \sum_{i \in F} \log(\sigma_i \pi) + \sum_{i \in C} \log \left(1 - \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu_i \sinh(z_i)] \right\}^{\tau_i} \right). \end{aligned} \quad (3.17)$$

The MLE $\hat{\boldsymbol{\theta}}$ of the vector $\boldsymbol{\theta}^T = (\mu, \sigma, \nu, \tau)$ of unknown parameters can be evaluated by maximizing the log-likelihood (3.17) numerically in the GAMLSS package of the R software. The advantage of using this package is that we can adopt many maximization methods, which will depend only on the current fitted model. When there are no explanatory variables or censored observations, we can use the `gamlssML` function for fitting (3.17) using a non-linear maximization algorithm. In the presence of 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 procedures used in the presence of censored data are the generalizations of the Rigby and Stasinopoulos (RS) and Cole and Green (CG) algorithms. All methods and algorithms are described by Rigby and Stasinopoulos (2005) and Stasinopoulos and Rigby (2007) and available in the GAMLSS package. The RS algorithm requires the first order derivatives of the logarithm of the density function (3.5) given in the above equations, and the second order derivatives. The RS method, different from the CG algorithm, does not use the cross derivatives, and thus it is faster for larger data sets.

An important consideration in the statistical analysis in 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. In particular, we now consider the test of homogeneity of variances for the ESC regression model based on the asymptotic distribution of the parameters. Under standard regularity conditions, the asymptotic distribution of $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is $N_p(\mathbf{0}, I(\boldsymbol{\theta})^{-1})$, where $I(\boldsymbol{\theta})$ is the expected information matrix. The multivariate normal $N_p(\mathbf{0}, \check{\mathbf{L}}(\hat{\boldsymbol{\theta}})^{-1})$ distribution can be used to construct approximate confidence intervals for the individual parameters, where $\check{\mathbf{L}}(\hat{\boldsymbol{\theta}})$ is the observed information matrix. Following (??), we generalize the scale parameter σ as $\boldsymbol{\sigma} = g_2(\mathbf{X}_2 \boldsymbol{\beta}_2)$, where $\mathbf{X}_{i2}$ is a matrix of explanatory variable values. For example, consider a matrix $\mathbf{X}_2$ ($n \times 2$) with the first column of ones corresponding to β_{02} , and the second column with the values of x_1 corresponding to β_{12} . We can test the homogeneity of variances between the levels (or ranges) of x_1 by testing the hypotheses $\mathcal{H}_0 : \beta_{12} = 0$

against $\mathcal{H}_a : \beta_{12} \neq 0$, where the Wald statistic is given by $T = \hat{\beta}_{12} / \sqrt{\hat{\mathbf{L}}(\hat{\boldsymbol{\theta}})_{\beta_{12}}^{-1}} \sim t_{(n-p-1)}$, and $\hat{\mathbf{L}}(\hat{\boldsymbol{\theta}})_{\beta_{12}}^{-1}$ is the $(p_1 + 2, p_1 + 2)$ element of the observed information matrix. Analogously, we can provide the same tests of hypotheses for the parameters $\boldsymbol{\mu}$, ν and τ .

3.4 Simulation Study

We conduct two Monte Carlo simulation studies to assess the finite sample behavior of the MLEs of the parameters for different sample sizes “ n ” and censoring percentages “ κ ”. In the first simulation, we consider the location model in (3.15), where $\mu_i = \beta_{01} + \beta_{11}x_i$, $\sigma_i = \sigma$, $\nu_i = \nu$ and $\tau_i = \tau$. In the second simulation, we consider the GAMLSS model in (3.15) by modeling the parameters using the explanatory variable x_i , namely: $\mu_i = \beta_{01} + \beta_{11}x_i$, $\sigma_i = \exp(\beta_{02} + \beta_{12}x_i)$, $\nu_i = \exp(\beta_{03} + \beta_{13}x_i)$ and $\tau_i = \exp(\beta_{04} + \beta_{14}x_i)$.

In the two simulations, the sample sizes are generated by taking $n = 50$ and 100 . The log-lifetimes denoted by $\log(T_1), \dots, \log(T_n)$ are generated from the ESC distribution using the qf (3.9), where the parameter vectors were fixed and evaluated using the explanatory variable x_i generated from a uniform $(0, 1)$ distribution. The censoring times, denoted by $C_1, \dots, C_n$, are randomly generated for censoring percentages $\kappa = 0.0, 0.1$ and 0.3 , respectively.

The lifetimes considered in each fit are evaluated as $\min[\log(C_i), \log(T_i)]$. For each configuration of n and κ , all results are obtained from 2,000 Monte Carlo replications and the simulations are carried out using the R programming language. For each replication, a random sample of size n is drawn from the ESC regression model (3.15) for survival censored data and the optim algorithm is used for maximizing the total log-likelihood function (3.17).

3.4.1 Location simulation

For the location model, the true parameter values used in the data-generating process are $\mu_i = 1 + 3x_i$, $\sigma = 3$, $\nu = 0.2$ and $\tau = 2$. For each fit, the average estimates (AEs), biases and means squared errors (MSEs) are evaluated. The results are given in Table 3.1.

Table 3.1. The AEs, biases and MSEs based on 2,000 simulations for the location ESC regression model when $\beta_{01}=1$, $\beta_{11}=3$, $\sigma = 3$, $\nu = 0.2$ and $\tau = 2$, for $n=50$ and 100 for censoring percentages $\kappa = 0.0, 0.1$ and 0.3 .

κ	Parameter	$n = 50$			Parameter	$n = 100$		
		AE	Bias	MSE		AE	Bias	MSE
0.0	β_0	1.326	0.326	2.920	β_0	1.185	0.185	1.033
	β_1	2.978	-0.022	5.968	β_1	3.044	0.044	2.117
	σ	2.628	-0.372	0.280	σ	2.704	-0.296	0.152
	ν	0.164	-0.036	0.007	ν	0.171	-0.029	0.003
	τ	2.053	0.053	0.200	τ	2.063	0.063	0.102
0.1	β_0	1.324	0.324	3.365	β_0	1.039	0.039	1.248
	β_1	3.036	0.036	6.402	β_1	3.414	0.414	3.029
	σ	2.732	-0.268	0.222	σ	2.817	-0.183	0.123
	ν	0.169	-0.031	0.006	ν	0.174	-0.026	0.003
	τ	2.187	0.187	0.269	τ	2.188	0.188	0.143
0.3	β_0	2.511	1.511	7.315	β_0	0.986	-0.014	1.564
	β_1	1.111	-1.889	12.032	β_1	3.450	0.450	3.121
	σ	3.024	0.024	0.332	σ	3.142	0.142	0.185
	ν	0.189	-0.011	0.024	ν	0.194	-0.006	0.004
	τ	2.553	0.553	1.590	τ	2.523	0.523	0.429

The estimated survival functions are displayed in Figure 3.4 by considering the AEs given in Table 3.1 for $n = 100$, and considering the maximum and minimum values of the generated x_i variable.

Figure 3.4. Some ESC survival functions at the true parameter values and at the AEs obtained in Table 3.1, considering $n = 100$ for the maximum and minimum of x_i when: (a) $\kappa=0$; (b) $\kappa=0.1$; (c) $\kappa=0.3$.

3.4.2 GAMLSS simulation

For the GAMLSS, the true parameter values used in the data-generating process are $\mu_i = 0.5 + 6x_i$, $\sigma_i = \exp(1.5 + 0.6x_i)$, $\nu_i = \exp(-3.5 + 3x_i)$ and $\tau_i = \exp(0.2 + 0.9x_i)$. For each fit, the AEs, biases and MSEs are reported in Table 3.2.

Table 3.2. The AEs, biases and MSEs based on 2,000 simulations of the ESC regression model when $\beta_{01}=0.5$, $\beta_{11}=6$, $\beta_{02} = 1.5$, $\beta_{12} = 0.6$, $\beta_{03} = -3.5$, $\beta_{13} = 3$, $\beta_{04} = 0.2$ and $\beta_{14} = 0.9$, for $n=50$ and 100 and under censoring percentages $\kappa = 0.0$, 0.1 and 0.3.

κ	Parameter	$n = 50$			Parameter	$n = 100$		
		AE	Bias	MSE		AE	Bias	MSE
0.0	β_{01}	0.547	0.047	5.845	β_{01}	0.471	-0.029	2.647
	β_{11}	7.041	1.041	29.142	β_{11}	6.756	0.756	13.629
	β_{02}	1.375	-0.125	0.072	β_{02}	1.414	-0.086	0.030
	β_{12}	0.587	-0.013	0.186	β_{12}	0.571	-0.029	0.089
	β_{03}	-4.058	-0.558	1.336	β_{03}	-3.861	-0.361	0.536
	β_{13}	3.490	0.490	2.414	β_{13}	3.273	0.273	1.061
	β_{04}	0.220	0.020	0.135	β_{04}	0.228	0.028	0.061
	β_{14}	0.908	0.008	0.456	β_{14}	0.895	-0.005	0.211
0.1	β_{01}	0.505	0.005	5.676	β_{01}	0.546	0.046	2.632
	β_{11}	6.903	0.903	28.215	β_{11}	6.664	0.664	16.902
	β_{02}	1.388	-0.112	0.064	β_{02}	1.446	-0.054	0.025
	β_{12}	0.656	0.056	0.218	β_{12}	0.597	-0.003	0.098
	β_{03}	-4.018	-0.518	1.171	β_{03}	-3.797	-0.297	0.457
	β_{13}	3.479	0.479	2.578	β_{13}	3.248	0.248	0.969
	β_{04}	0.265	0.065	0.132	β_{04}	0.309	0.109	0.063
	β_{14}	0.975	0.075	0.494	β_{14}	0.865	-0.035	0.211
0.3	β_{01}	0.889	0.389	7.340	β_{01}	0.636	0.136	3.020
	β_{11}	6.381	0.381	21.376	β_{11}	6.319	0.319	9.264
	β_{02}	1.450	-0.050	0.092	β_{02}	1.482	-0.018	0.040
	β_{12}	0.718	0.118	0.307	β_{12}	0.753	0.153	0.183
	β_{03}	-3.939	-0.439	1.576	β_{03}	-3.807	-0.307	0.640
	β_{13}	3.499	0.499	3.137	β_{13}	3.478	0.478	1.580
	β_{04}	0.510	0.310	0.272	β_{04}	0.508	0.308	0.155
	β_{14}	0.789	-0.111	0.511	β_{14}	0.790	-0.110	0.206

The estimated survival functions are displayed in Figure 3.5 and the AEs are listed in Table 3.2 for $n = 100$, and considering the maximum and minimum values of the generated x_i variable.

The results of the Monte Carlo study in Tables 3.1 and 3.2 indicate that the MSEs of the MLEs of the parameters decay toward zero when the sample size increases, as expected under first-order asymptotic theory. Note that the results of the GAMLSS simulation, presented in Table 3.2, should be

Figure 3.5. Some ESC survival functions at the true parameter values and at the AEs obtained in Table 3.2, considering $n = 100$ for the maximum and minimum of x_i when: (a) $\kappa=0$; (b) $\kappa=0.1$; (c) $\kappa=0.3$.

interpreted by peers due to the fit of β_{ik} influences the fit of β_{jk} . If n increases, the AEs tend to be closer to the true parameter values. This fact supports that the asymptotic normal distribution provides an adequate approximation to the finite sample distribution of the MLEs. The normal approximation can oftentimes be improved by using bias adjustments to these estimators. In general, for the ESC regression models, the variances and MSEs increase when the censoring percentage increases. This fact can be noted in Figures 3.4 and 3.5.

3.5 Study of model misspecification

To assess the behavior of the MLEs of the parameters in the ESC regression model when it is poorly specified, we carry out a Monte Carlo simulation study based on 1,000 replications using the GAMLSS. The logarithms of the lifetime data are generated from the log-Weibull (y, μ, σ) and normal (y, μ, σ) heteroscedastic regression models (traditional models used in the survival analysis) for selected parameters $\mu = \beta_{01} + \beta_{11} x_1$ and $\sigma = \exp(\beta_{02} + \beta_{12} x_1)$, where the covariate x_i is generated from a binomial $(n, 0.5)$ distribution. The censored indicators are generated randomly by fixing the censoring percentage. We consider the configuration with sample size $n = 100$, $\beta_{01} = 4.5$, $\beta_{11} = 1.5$, $\beta_{02} = -1.5$, $\beta_{12} = 1.5$ and censoring percentages of $\rho = 0\%$, 10% and 30% to generate the samples. We fit the ESC regression model to each generated data set. The results of this study are given in Table 3.3, where we can note that an increasing in censoring percentage in general implies an increasing in the MSEs. There is a small sample bias in the estimation of the parameters of this regression model. Hence, it can provide consistent MLEs even when the data are generated from a different model.

Table 3.3. Mean estimates and MSEs (in parentheses) of the MLEs of the parameters in the log-Weibull and normal heteroscedastic regression models.

θ	log-Weibull			normal		
	$\rho = 0\%$	$\rho = 10\%$	$\rho = 30\%$	$\rho = 0\%$	$\rho = 10\%$	$\rho = 30\%$
β_{01}	4.510(0.005)	4.526(0.006)	4.553(0.009)	4.452(0.006)	4.467(0.006)	44.488(0.006)
β_{11}	1.569(0.087)	1.611(0.101)	1.701(0.123)	1.385(0.076)	1.427(0.079)	1.482(0.080)
β_{02}	-1.905(0.224)	-1.838(0.275)	-1.734(0.209)	-1.744(0.086)	-1.687(0.072)	-1.545(0.025)
β_{12}	1.498(0.041)	1.494(0.043)	1.514(0.047)	1.496(0.029)	1.505(0.029)	1.503(0.033)
ν	1.207(-)	1.271(-)	1.280(-)	1.084(-)	1.122(-)	1.150(-)
τ	0.608(-)	0.637(-)	0.733(-)	1.309(-)	1.339(-)	1.490(-)

3.6 Sensitivity and residual analysis

Since regression models are sensitive to the underlying model assumptions, performing a sensitivity analysis is strongly advisable. Cook (1986) used this idea to motivate the assessment of influence analysis. He suggested 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 (Cook and Weisberg, 1982), in which the effects of completely removing cases from the analysis are studied.

3.6.1 Global influence

A first tool to perform sensitivity analyses, as stated before, is by means of global influence starting from case-deletion. Case-deletion is a common approach to study the effect of dropping the i th case from the data set. The case-deletion model for model (3.15) is given by

$$y_l = \mu_l + \sigma_l z_l, \quad l = 1, \dots, n, \quad l \neq i, \quad (3.18)$$

where the random error Z_l has a density function $f(z_l; \nu_l, \tau_l)$ given in (3.7). Of course, not always the explanatory variables will be modeling all parameters. For example, if we consider the class of location in (3.18), the case-deletion model reduces to

$$y_l = \mu_l + \sigma z_l, \quad l = 1, \dots, n, \quad l \neq i,$$

where the random error Z_l has the density function $f(z_l; \nu, \tau)$.

In the following, a quantity with subscript “ (i) ” means the original quantity with the i th case deleted. For model (3.18), the log-likelihood function of θ is denoted by $l_{(i)}(\theta)$. Let $\hat{\theta}_{(i)}^T = (\hat{\mu}_{(i)}^T, \hat{\sigma}_{(i)}^T, \hat{\nu}_{(i)}^T, \hat{\tau}_{(i)}^T)$ be the MLE of μ, σ, ν and τ from $l_{(i)}(\theta)$. To assess the influence of the i th case on the MLE $\hat{\theta}^T = (\hat{\mu}^T, \hat{\sigma}^T, \hat{\nu}^T, \hat{\tau}^T)$, the basic idea is to compare the difference between $\hat{\theta}_{(i)}$ and $\hat{\theta}$. If deletion of a case seriously influences the estimates, for example changing the inference, more attention should be given to that case. Hence, if $\hat{\theta}_{(i)}$ is far from $\hat{\theta}$, then the i th case is regarded as an influential observation. A first measure of the global influence is defined as the standardized norm of $(\hat{\theta}_{(i)} - \hat{\theta})$, known as the generalized Cook distance, defined by

$$GD_i(\theta) = (\hat{\theta}_{(i)} - \hat{\theta})^T [-\ddot{\mathbf{L}}(\hat{\theta})](\hat{\theta}_{(i)} - \hat{\theta}).$$

Another alternative is to assess values $GD_i(\mu)$, $GD_i(\sigma)$, $GD_i(\nu)$ and $GD_i(\tau)$, which reveal the impact of the i th observation on the estimates of μ, σ, ν and τ , respectively. Another popular measure of the difference between $\hat{\theta}_{(i)}$ and $\hat{\theta}$ is the likelihood distance defined by

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

3.6.2 Local influence

Cook (1986) suggested to give weights to the observations instead of removing them. Local influence calculation can be carried out for model (3.15). If likelihood displacement $LD(\omega) = 2\{l(\hat{\theta}) - l(\hat{\theta}_\omega)\}$ is used, where $\hat{\theta}_\omega$ denotes the MLE under the perturbed model, the normal curvature for θ in the direction $\mathbf{d}$, $\|\mathbf{d}\| = 1$, is given by $C_{\mathbf{d}}(\theta) = 2|\mathbf{d}^T \Delta^T \ddot{\mathbf{L}}_{\theta\theta}^{-1} \Delta \mathbf{d}|$, where Δ is a $p \times n$ matrix that depends on the perturbation scheme, whose elements are given by $\Delta_{vi} = \partial^2 l(\theta|\omega) / \partial \theta_v \partial \omega_i$, $i = 1, \dots, n$ and $v = 1, \dots, p$, evaluated at $\hat{\theta}$ and ω_0 , and ω_0 is the no perturbation vector. We can also calculate normal curvatures $C_{\mathbf{d}}(\mu)$, $C_{\mathbf{d}}(\sigma)$, $C_{\mathbf{d}}(\nu)$ and $C_{\mathbf{d}}(\tau)$ to perform various index plots, for instance, the index plot of $\mathbf{d}_{max}$, the eigenvector corresponding to $C_{\mathbf{d}_{max}}$, the largest eigenvalue of the matrix $\mathbf{B} = -\Delta^T \ddot{\mathbf{L}}_{\theta\theta}^{-1} \Delta$ and the index plots of $C_{\mathbf{d}_i}(\mu)$, $C_{\mathbf{d}_i}(\sigma)$, $C_{\mathbf{d}_i}(\nu)$ and $C_{\mathbf{d}_i}(\tau)$, named the total local influence (Lesaffre and Verbeke,

1998), where $\mathbf{d}_i$ denotes an $n \times 1$ vector of zeros with one at the i th position. Thus, the curvature in the direction $\mathbf{d}_i$ takes the form $C_i = 2|\Delta_i^T \ddot{\mathbf{L}}_{\theta\theta}^{-1} \Delta_i|$, where Δ_i^T denotes the i th row of Δ . It is usual to point out those cases such that $C_i \geq 2\bar{C}$, where $\bar{C} = \frac{1}{n} \sum_{i=1}^n C_i$. In some situations, the information of the matrix $\mathbf{B}$ may be contained not only in the first eigenvalue, then an alternative influence measure for the i th observation is $U_i = \sum_{k=1}^{n_1} \lambda_k e_{ki}^2$, where $\{(\lambda_k, \mathbf{e}_k) | k = 1, \dots, n\}$ are the eigenvalue-eigenvector pairs of $\mathbf{B}$ with $\lambda_1 \geq \dots \geq \lambda_{n_1} \geq \lambda_{n_1+1} = \dots = \lambda_n = 0$ and $\{\mathbf{e}_k = (e_{k1}, \dots, e_{kn})^T\}$ is the associated orthonormal basis. Zhu *et al.* (2007) studied the influence measure u_i systematically under a case-weight perturbation. Thus, this influence measure expresses local sensitivity to the log-likelihood of the perturbations.

Next, we obtain under model (3.15) and log-likelihood function (3.17), for three perturbation schemes, the matrix

$$\Delta = (\Delta_{vi})_{p \times n} = \left(\frac{\partial^2 l(\theta|\omega)}{\partial \theta_v \partial \omega_i} \right)_{p \times n}, \quad v = 1, \dots, p \quad \text{and} \quad i = 1, \dots, n.$$

3.6.2.1 Case-weight perturbation

Consider the vector of weights $\omega = (\omega_1, \dots, \omega_n)^T$, where $0 \leq \omega_i \leq 1$. A perturbed log-likelihood function, allowing different weights for different observations, can be defined in the form $l(\theta|\omega) = \sum_{i \in F} \omega_i \log f(y_i) + \sum_{i \in C} \omega_i \log S(y_i)$. Also, let $w_0 = (1, \dots, 1)^T$ be the vector of no perturbation such that $l(\theta|w_0) = l(\theta)$. In this case, the log-likelihood function takes the form

$$l(\theta|\omega) = \sum_{i \in F} \omega_i \left[-\log d_i + (\tau_i - 1) \log(h_i) + \log \cosh(z_i) + \log(\tau_i \nu_i) - \log(\sigma_i \pi) \right] + \sum_{i \in C} \omega_i \log [1 - h_i^{\tau_i}],$$

where $h_i = 0.5 + \pi^{-1} \arctan[\nu_i \sinh(z_i)]$ and $d_i = [1 + \nu_i^2 \sinh^2(z_i)]$. The matrix $\Delta = (\Delta_{\mu}^T, \Delta_{\sigma}^T, \Delta_{\nu}^T, \Delta_{\tau}^T)^T$ can be calculated numerically.

3.6.2.2 Response perturbation

Since the values of y_i have different variances, they require a scaling of the perturbation vector ω by an estimator of the standard deviation of y_i . We shall consider that each y_i is perturbed as $y_{iw} = y_i + \omega_i S_y$, where S_y is a scale factor that may be estimated by the standard deviation of y and $\omega_i \in \mathbf{R}$. Then, the perturbed log-likelihood function becomes

$$l(\theta) = \sum_{i \in F} \left[-\log d_i^* + \log \cosh(z_i^*) + (\tau_i - 1) \log(h_i^*) + \log(\tau_i \nu_i) - \log(\sigma_i \pi) \right] + \sum_{i \in C} \log(1 - h_i^{*\tau_i}),$$

where $h_i^* = 0.5 + \pi^{-1} \arctan[\nu_i \sinh(z_i^*)]$, $d_i^* = [1 + \nu_i^2 \sinh^2(z_i^*)]$ and $z_i^* = (y_i + \omega_i S_y - \mu_i)/\sigma_i$. The matrix $\Delta = (\Delta_{\mu}^T, \Delta_{\sigma}^T, \Delta_{\nu}^T, \Delta_{\tau}^T)^T$ can be calculated numerically.

3.6.2.3 Explanatory variable perturbation

We consider an additive perturbation on a particular continuous explanatory variable, namely $x_1[i, t]$, by setting $x_1[i, t\omega] = x_1[i, t] + \omega_i S_x$, where S_x is a scaled factor, $\omega_i \in \mathbf{R}$. Note that the explanatory variable $x_1[i, t]$ is related only to the location parameter μ . However, this perturbation scheme can be extended by considering different numbers of explanatory variables for different parameters.

This perturbation scheme leads to the perturbed log-likelihood function

$$l(\theta) = \sum_{i \in F} \left[-\log d_i^* + \log \cosh(z_i^*) + (\tau_i - 1) \log(h_i^*) + \log(\tau_i \nu_i) - \log(\sigma_i \pi) \right] + \sum_{i \in C} \log(1 - h_i^{*\tau_i}),$$

where $h_i^* = 0.5 + \pi^{-1} \arctan[\nu_i \sinh(z_i^*)]$, $d_i^* = [1 + \nu_i^2 \sinh^2(z_i^*)]$, $z_i^* = (y_i - \mu_i^*)/\sigma_i$ and $\mu_i^* = \beta_{01} + \beta_{11}x_1[i, 2], \dots, \beta_{t1}(x_1[i, t] + \omega_i S_x \dots, \beta_{p1}x_1[p_1, 1])$. The matrix $\Delta = (\Delta_{\mu}^T, \Delta_{\sigma}^T, \Delta_{\nu}^T, \Delta_{\tau}^T)^T$ can be calculated numerically.

3.6.3 Residual Analysis

In order to study departures from the error assumption and the presence of outliers, we consider the martingale residual proposed by Barlow and Prentice (1998) and the transformation of this residual. More details may be found in Ortega *et al.* (2003).

The martingale residuals, recommended in counting processes, are defined by $r_{M_i} = \delta_i + \log[S(y_i; \hat{\beta})]$, where $\delta_i = 0, 1$ denotes a censored and uncensored observation, respectively, and $S(y_i; \hat{\beta})$ denotes the survival function of Y discussed in Section 3.1. Recently, several authors have studied the martingale residual for some regression models. Silva *et al.* (2008) proposed using the martingale residual for the log-Burr XII regression model considering censored data, Cancho *et al.* (2009) studied the residuals for the log-exponentiated-Weibull regression model with cure rate, Ortega *et al.* (2014) derived the martingale residual for the odd Weibull regression models for censored data, among others.

This residual was introduced in the counting process (Fleming and Harrington, 1991) and can be expressed in the ESC regression models as

$$r_{M_i} = \begin{cases} 1 - \hat{\tau}_i \log(2\pi) + \log \left[(2\pi)^{\hat{\tau}_i} - \left\{ \pi + 2 \arctan[\hat{\nu}_i \sinh(\hat{z}_i)] \right\}^{\hat{\tau}_i} \right] & \text{if } i \in F \\ -\hat{\tau}_i \log(2\pi) + \log \left[(2\pi)^{\hat{\tau}_i} - \left\{ \pi + 2 \arctan[\hat{\nu}_i \sinh(\hat{z}_i)] \right\}^{\hat{\tau}_i} \right] & \text{if } i \in C, \end{cases} \quad (3.19)$$

where $\hat{z}_i = (y_i - \hat{\mu}_i)/\hat{\sigma}_i$, $\mu_i = \hat{\beta}_{01} + \dots + x_1[i, p_1 + 1]\hat{\beta}_{p_{11}}$, $\sigma_i = \exp(\hat{\beta}_{02} + \dots + x_2[i, p_2 + 1]\hat{\beta}_{p_{22}})$, $\nu_i = \exp(\hat{\beta}_{03} + \dots + x_3[i, p_3 + 1]\hat{\beta}_{p_{33}})$ and $\tau_i = \exp(\hat{\beta}_{04} + \dots + x_4[i, p_4 + 1]\hat{\beta}_{p_{44}})$. In fact, r_{M_i} ranges from a maximum value $+1$ and minimum value $-\infty$. A disadvantage of the martingale residual is that the distribution of r_{M_i} is markedly skewed, and so it fails to have similar properties to those of the normal distribution. Suitable transformations to achieve a more normal shaped form would be more appropriate for residual analysis.

Another possibility is to use a transformation of the martingale residual based on the deviance residuals for the Cox model in the case of no time-dependent covariates (Therneau *et al.*, 1990). We shall use this transformation of the martingale residual in order to have a new residual symmetrically distributed around zero. A more extensive examination of this residual is given by Leiva *et al.* (2007) and Ortega *et al.* (2008). Thus, a martingale-type residual for the ESC regression model can be expressed as

$$r_{D_i} = \text{sign}(r_{M_i}) \left\{ -2 \left[r_{M_i} + \delta_i \log(\delta_i - r_{M_i}) \right] \right\}^{1/2},$$

where r_{M_i} is defined in equation (3.19) for $i \in F$ ($\delta_i = 1$) or $i \in C$ ($\delta_i = 0$).

For uncensored data, we can use the diagnostic tools in the gamlss package. The first technique consists in the normalized randomized quantile residuals (Dunn and Smyth, 1996) given by $\hat{r}_i = \Phi^{-1}(u_i)$, where $\Phi^{-1}(\cdot)$ is the inverse cdf of a standard normal variate and $u_i = F(y_i | \hat{\theta}_i)$.

The second technique already known in the literature is the normal probability plot with envelope. Atkinson (1985) suggested the construction of envelopes to enable better interpretation of the normal probability plots of the residuals. Such envelopes are simulated confidence bands that contain the residuals, such that if the model is well fitted, the majority of points will be within these bands and randomly distributed. The construction of the confidence bands follows the steps:

- Fit the proposed model, we evaluate the normalized randomized quantile residuals $\hat{r}_i$;
- Simulate k samples of size n of the response variable using the fitted model;
- For each sample, we compute the residuals $\hat{r}_{ij}$, $j = 1, 2, \dots, k$ and $i = 1, 2, \dots, n$;
- Arrange each group of n residuals in rising order to obtain $\hat{r}_{(i)j}$;

- For each i , obtain the minimum and maximum $\hat{r}_{(i)j}$, namely:

$$r_{(i)I} = \min\{r_{(i)j} : 1 \leq j \leq k\} \quad \text{and} \quad r_{(i)S} = \max\{r_{(i)j} : 1 \leq j \leq k\};$$

- Include the minimum and maximum together with the values of $\hat{r}_i$ against the expected percentiles of the standard normal distribution.

The minimum and maximum values of $\hat{r}_{(i)j}$ define the envelope. If the model under study is correct, the observed values of $\hat{r}_i$ should be inside the bands and distributed randomly.

3.7 Applications

In this section, we provide two applications to real data to illustrate the flexibility of the ESC regression model. The computations are performed using the `gamlss` subroutine in the R software and the script is described in the Appendix. For the first data set, we prove empirically the flexibility of the new regression model when all parameters are modeled by explanatory variables (complete model). For the second data set, we present an application, where the scale and skewness parameters are modeled by explanatory variables. For both applications we provide the goodness-of-fit statistics Akaike information criterion (AIC) and Bayesian information criterion (BIC). The computational codes for the applications in sections 3.7.1 and 3.7.2 are available on the Web at <http://goo.gl/zANZuz> and <http://goo.gl/ZBf8R8>, respectively.

3.7.1 Shrimp data

Consider the data on biometric measurements in shrimps of *farfantepenaeus brasiliensis* species. These data were obtained from three regions of the Rio Grande do Norte state in Brazil, for which the objective was to relate the weights of the shrimps in each region. The importance of characterizing the weights of shrimps per region is discussed by Pinheiro (2008).

To exemplify the new propose, we consider the full sample ($n = 120$), where the response variable t_i represents the i th shrimp weight in grams and the three groups of region are defined by dummy variables: *Baia formosa* ($x_{i1} = 0$ and $x_{i2} = 0$), *Diogo Lopes* ($x_{i1} = 1$ and $x_{i2} = 0$) and *Touros* ($x_{i1} = 0$ and $x_{i2} = 1$). Let the random variable $y_i = \log(t_i)$ have the ESC distribution (3.5). As a preliminary analysis, we note that the explanatory variable *region* affects the location, scale, bimodality and asymmetry parameters. This fact can easily be observed in Figure 3.6.

Figure 3.6. The empirical density of Y in the different regions.

Next, we present results by fitting the model

$$y_i = \mu_i + \sigma_i z_i,$$

where z_i has density function (3.7) and the model parameters are defined by

$$\begin{aligned}\mu_i &= \beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2}, & \sigma_i &= \exp(\beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2}), \\ \nu_i &= \exp(\beta_{03} + \beta_{13}x_{i1} + \beta_{23}x_{i2}) \quad \text{and} \quad \tau_i &= \exp(\beta_{04} + \beta_{14}x_{i1} + \beta_{24}x_{i2}).\end{aligned}$$

Table 3.4 provides the MLEs, their approximate standard errors and p -values, all quantities obtained from the fitted ESC regression model. The values of the goodness-of-fit statistics are $AIC = 142.9$ and $BIC = 176.3$. The results in Table 3.4 reveal that the explanatory variable *region* should be used to model the location, scale, bimodality and skewness parameters at the 5% level. Therefore, we can conclude that for each region, the weights of shrimps have different forms (bimodal and unimodal), different location scales and asymmetry, and then they can not be fitted only with a location model.

Table 3.4. MLEs of the parameters and their approximate standard errors from the fitted ESC regression model to the shrimp data.

Parameter	Estimate	SE	p -value	Parameter	Estimate	SE	p -value
β_{01}	2.721	0.034	<0.001	β_{03}	-2.616	0.613	<0.001
β_{11}	-1.163	0.398	0.004	β_{13}	2.059	0.777	0.009
β_{21}	0.594	0.091	<0.001	β_{23}	2.425	0.754	0.001
β_{02}	-2.235	0.175	<0.001	β_{04}	-0.189	0.232	0.416
β_{12}	1.223	0.387	0.002	β_{14}	0.655	0.713	0.360
β_{22}	-0.057	0.495	0.908	β_{24}	-1.165	0.595	0.052

3.7.1.1 Global influence analysis

Here, we compute the case deletion measures $GD_i(\theta)$ and $LD_i(\theta)$ for the shrimp data. The results of such influence measure index plots are displayed in Figure 3.7. We may note that the 62th observation is a possible influential observation.

Figure 3.7. Index plots for θ : (a) $GD_i(\theta)$ (Generalized Cook's Distance) and (b) $LD_i(\theta)$ (Likelihood Distance).

3.7.1.2 Local influence analysis

In this section, we perform the local influence analysis for the shrimp data using the ESC regression model.

Case-weight perturbation

By applying the local influence methodology, where the case-weight perturbation is used, the four largest eigenvalues of the matrix $\mathbf{B}$ are 1.65, 1.64, 1.26 and 1.12. Figure 3.8 displays the index plots

of the U_i measure and the total influence C_i . These plots reveal that the 62th observation also appears as possible influential observation.

Figure 3.8. Index plots for θ (case-weight perturbation): (a) d_{max} and (b) total local influence.

Response perturbation

Next, the influence of perturbations in the observed times is analyzed. Here, we adopt the U_i measure instead of d_{max} because the first eight eigenvalues are large. Figure 3.9 displays the index plot of the U_i measure and the total local influence C_i .

Figure 3.9. Index plots for θ (response perturbation): (a) d_{max} and (b) total local influence.

Under the sensitivity analysis, we note that the 62th observation once more appears as a possible influential point. In fact, this shrimp has the largest weight for Diogo Lopes region, being very different from the other measurements. The shrimps detected as possible influential observations in Figure 3.9 represent the measurements $y_{105} = 2.89$ and $y_{107} = 2.88$ of the Touro region. Combining with the plots of Figure 3.6, we can note that these two shrimps stabilize the growth of the density.

3.7.1.3 Residual analysis

In order to detect possible outlying observations as well as departures from the assumptions made for the ESC regression model, we present in Figure 3.10 the index plot as well as the normal probability plot with generated confidence band for the quantile residual. Note that the quantile residual seems to follow approximately a normal distribution, thus indicating a suitable fitted model. Note that the observations detected in the influence analysis are not detected in the residual analysis.

In order to assess whether the model fits the data appropriately, the empirical cdf and estimated cdf of the ESC regression model are plotted in Figure 3.11 for different regions. We conclude that the Exp-ESC regression model provides a very good fit to the shrimp data.

Figure 3.10. (a) Index plot of the quantile residuals for the shrimp data. (b) Normal probability plot with envelope for the quantile residuals from the fitted ESC regression model to the shrimp data.

Figure 3.11. Estimated cumulative fitted values from the ESC fitted model to the shrimp data.

3.7.2 Entomology data

In the second application, we take a data set from a study carried out at the Department of Entomology of the Luiz de Queiroz School of Agriculture, University of S o Paulo. Such study aims to assess the longevity of the mediterranean fruit fly (*ceratitis capitata*), which is considered a pest in agriculture. Instead of using an insecticide, Silva *et al.* (2013) conducted a study using small portions of food containing substances extracted from a tree called “neem”. The experiment was completely randomized with eleven treatments, consisting of different extracts of the neem tree at concentrations of 39, 225 and 888 ppm, where the response variable is the lifetime of the adult flies in days after exposure to the treatments. The experimental period was set at 51 days, so that the numbers of larvae that survived beyond this period are considered as censored observations. From the results of the experiment, these eleven treatments are allocated into two groups, namely:

- Group 1: Control 1 (deionized water); Control 2 (acetone - 5%); aqueous extract of seeds (AES) (39 ppm); AES (225 ppm); AES (888 ppm); methanol extract of leaves (MEL) (225 ppm); MEL (888 ppm); and dichloromethane extract of branches (DMB) (39 ppm).
- Group 2: MEL (39 ppm); DMB (225ppm) and DMB (888 ppm).

Let t_i be the lifetime of *ceratitis capitata* adults in days, δ_i the censoring indicator and x_{i1} the dummy variable indicating the groups (0=group 1 and 1=group 2). In a preliminary analysis, we note that only the scale and skewness parameters require explanatory variables. Next, we present results by

fitting the model

$$y_i = \beta_{01} + \sigma_i z_i,$$

where z_i , for $i = 1, \dots, 172$, has density function $f(z_i; \nu, \tau_i)$ given by (3.7) and the model parameters are given by

$$\mu_i = \beta_{01}, \quad \sigma_i = \exp(\beta_{02} + \beta_{12}x_{i1}), \quad \nu_i = \exp(\beta_{03}) \quad \text{and} \quad \tau_i = \exp(\beta_{04} + \beta_{14}x_{i1}).$$

Table 3.5 provides the MLEs, their approximate standard errors and p -values obtained from the fitted ESC regression model. We can conclude that the explanatory variable *group* should be used to model the scale and skewness parameters at the 1% level. The goodness-of-fit statistics obtained are $AIC = 309.3$ and $BIC = 328.2$. Recently, Cordeiro *et al.* (2015) fitted the log-generalized Weibull-log-logistic (LGW-LL) to these data and obtained the statistics $AIC = 341$ and $BIC = 357$. We conclude that the ESC regression model provides a good fit to these data.

Table 3.5. MLEs of the parameters and their approximate standard errors from the fitted ESC regression model to the entomology data.

Parameter	Estimate	SE	p -value	Parameter	Estimate	SE	p -value
β_{01}	3.013	0.024	<0.001	β_{03}	1.218	0.112	<0.001
β_{02}	-0.012	0.119	0.913	β_{04}	0.100	0.085	0.242
β_{12}	-0.895	0.234	<0.001	β_{14}	-0.893	0.175	<0.001

3.7.2.1 Global influence analysis

Here, we compute the case deletion measures $GD_i(\theta)$ and $LD_i(\theta)$ for the entomology data. The results of such influence measure index plots are displayed in Figure 3.12. Based on these plots, we note that the cases 92 and 133 are possibly influential observations.

Figure 3.12. Index plots for θ : (a) $GD_i(\theta)$ (Generalized Cook's Distance) and (b) $LD_i(\theta)$ (Likelihood Distance).

3.7.2.2 Local influence analysis

Case-weight perturbation

By applying the local influence methodology, where case-weight perturbation is applied, we obtain $C_{\mathbf{d}_{max}} = 1.15$ as the maximum curvature. Figure 3.13 display the index plots of the eigenvector corresponding to $\mathbf{d}_{max}$ and the total influence C_i . We may conclude that the observations 145 and 157 present larger influence.

Figure 3.13. Index plots for θ (case-weight perturbation): (a) d_{max} and (b) total local influence.

Response perturbation

The influence of perturbing the observed response Y will be analyzed. The value for the maximum curvature obtained is $C_{d_{max}} = 10.41$. Figure 3.14 display the index plots for d_{max} and total local influence C_i . We may conclude that the observations 96 and 153 are possible influential points.

Figure 3.14. Index plots for θ (response perturbation): (a) d_{max} and (b) total local influence.

The global influential analysis indicates that the observations 92 and 133 are possible influential. The 92th observation has the large lifetime of the group 2 and the 133th observation has the smallest lifetime of the group 1. Under the local influential analysis (case-weight perturbation), the observations 145 and 157 are detected and they represent the smallest lifetimes of the group 2 with lifetimes $t_{145} = t_{157} = 1$. Finally, with the local influential analysis (response perturbation), the detected observations 96th and 153th are the intermediary measures of the group 2.

3.7.2.3 Residual analysis

In order to detect possible outliers as well as departures from the assumptions made for the ESC regression model, we present in Figure 3.15 the normal probability plot with generated confidence band and the index plot for the martingale-type residual. By analyzing these plots, the asymmetry is observed. However, there is no indication of departures from the assumptions made for the model as well as the presence of outlying observations.

Finally, in order to assess if the model is appropriate, the empirical and estimated survival functions of the ESC regression model are plotted in Figure 3.16 for the different groups. We may conclude from the plots that the ESC regression model provides a suitable fit to the entomology data.

Figure 3.15. (a) Normal probability plot with envelope for the martingale-type residual r_{D_i} from the fitted ESC regression model to the entomology data. (b) Index plot of the martingale-type residual r_{D_i} for the entomology data.

Figure 3.16. Estimated and empirical survival functions for the entomology data.

3.8 Conclusions

In this paper, we propose a general class of exponentiated sinh Cauchy (ESC) regression models, where the mean, dispersion, skewness and bimodal parameters vary across observations through regression structures. The former class of regression models is very suitable for modeling censored and uncensored lifetime data. The proposed model serves as an important extension to several existing regression models and could be a valuable addition to the literature. We use the GAMLSS script in the R package to obtain the maximum likelihood estimates and perform asymptotic tests for the model parameters based on the asymptotic distribution of the estimates. We offer some interesting insights, especially regarding model checking, and provide applications of influence diagnostics (global, local and total influence) in the proposed class of regression models with censored data. We also discuss the adequacy of the regression models via martingale-type and quantile residuals. Several simulation studies are performed for different parameter settings, sample sizes and censoring percentages. Moreover, the usefulness of the model is also illustrated through the analysis of real data sets. Finally, the proposed algorithm for estimating the parameters in the probability density, cumulative distribution and quantile functions has been coded and implemented in the GAMLSS script available in the paper.

3.9 Script for the ESC regression model

Here, we provide a brief discussion of the script for the ESC regression model implemented in the GAMLSS R package. The first step to run the codes is load the `gamlss` and `gamlss.cens` packages as well as the ESC model codes. After loading the codes, the pdf, cdf and qf will be available to be used. It is also available the function to generate random values having the ESC distribution.

In the example below, we present two ways to obtain the MLEs of the model parameters for uncensored and censored data. For both models, `m1` and `m2`, we are modeling all parameters with the explanatory variable X . After fitting the selected models, we can access the goodness-of-fit statistics. Finally, the codes to access the residual analysis, for uncensored and censored, respectively, are reported.

```
library(gamlss); library(gamlss.cens); source("https://goo.gl/DxWFB6")
dESC(y,mu,sigma,nu,tau) #pdf
pESC(q,mu,sigma,nu,tau) #cdf
qESC(p,mu,sigma,nu,tau) #qf
rESC(n,mu,sigma,nu,tau) #sample
m1=gamlss(y~X, sigma.fo=~X, nu.fo=~X,tau.fo=~X,family="ESC")
m2=gamlss(Surv(y,delta)~X,sigma.fo=~X, nu.fo=~X,tau.fo=~X,family="ESC")
AIC(m1); BIC(m1)
#Residual analysis
plot(m1$residuals,ylim=c(-3,3),ylab="Quantile residuals")
rm=delta+log(1-pESC(y,m2$mu.fv,m2$sigma.fv,m2$nu.fv,m2$tau.fv))
rd=sign(rm)*(-2*(rm+log(delta-rm)))^(0.5)
plot(rd,ylab="Martingale-type residual",pch=16,ylim=c(-3,3))
```

References

- Atkinson, A.C. (1985). *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis*. Oxford: Clarendon Press.
- Barlow, W.E. and Prentice, R.L. (1988). Residuals for relative risk regression. *Biometrika*, **75**, 65–74.
- Cancho, V.G., Ortega, E.M.M. and Bolfarine, H. (2009). The log-exponentiated-Weibull regression models with cure rate: local influence and residual analysis. *Journal of Data Science*, **7**, 433–458.
- Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society*, **48**, 133–169.
- Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hill.
- Cooray, K. (2013). Exponentiated Sinh Cauchy Distribution with Applications. *Communications in Statistics-Theory and Methods*, **42**, 3838–3852.
- Cordeiro, G.M., Alizadeh, M., Ramires, T.G. and Ortega, E.M.M. (2016). The Generalized Odd Half-Cauchy Family of Distributions: Properties and Applications. *Communications in Statistics-Theory and Methods*. DOI:10.1080/03610926.2015.1109665.
- Cordeiro, G.M., Ortega, E.M.M. and Ramires, T.G. (2015). A new generalized Weibull family of distributions: mathematical properties and applications. *Journal of Statistical Distributions and Applications*, **2**, 1–25.
- Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.
- Fleming, T.R. and Harrington, D.P. (1991). *Counting processes and survival analysis*. John Wiley & Sons.

- Gradshteyn, I.S. and Ryzhik, I.M. (2007). *Table of Integrals, Series, and Products*, seventh edition. Academic Press, San Diego.
- Gupta, R.C., Gupta, P.L. and Gupta, R.D. (1998). Modeling failure time data by Lehman alternatives. *Communications in Statistics Theory and Methods*, **27**, 887–904.
- Gupta, R.D. and Kundu, D. (2001). Exponentiated exponential family: an alternative to Gamma and Weibull distributions. *Biometrical Journal*, **43**, 117–130.
- Hashimoto, E.M., Ortega, E.M.M., Cordeiro, G.M. and Barreto, M.L. (2012). The Log-Burr XII Regression Model for Grouped Survival Data. *Journal of biopharmaceutical statistics*, **22**, 141–159.
- Leiva, V., Barros, M., Paula, G.A., Galea, M. (2007). Influence diagnostics in log-Birnbaum-Saunders regression models with Censored Data. *Computational Statistics and Data Analysis*, **51**, 5694–5707.
- Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **54**, 570–582.
- Ortega, E.M.M., Bolfarine, H. and Paula, G.A. (2003). Influence diagnostics in generalized log-gamma regression models. *Computational Statistics and Data Analysis*, **42**, 165–186.
- Ortega, E.M.M., Cordeiro, G.M., Lemonte, A.J. and Cruz, J.N. (2016). The Log-Odd Birbaum-Sauders Regression Model. *Journal of Testing and Evaluation*, (Submited).
- 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.
- 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. and Kattan, M.W. (2013). The log-beta Weibull regression model with application to predict recurrence of prostate cancer. *Statistical Papers*, **54**, 113–132.
- Ortega, E.M.M., Paula, G.A. and Bolfarine, H. (2008). Deviance residuals in generalized log-gamma regression models with censored observations. *Journal of Statistical Computation and Simulation*, **78**, 747–764.
- Pinheiro, A.P. (2008). Caracterização genética e biométrica das populações de camarão rosa *Farfantepenaeus brasiliensis* de três localidades da costa do Rio Grande do Norte. *Ecology and Natural Resources, Federal Univ. of São Carlos, SP, Brazil*.
- R Core team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
- Ramires, T.G., Ortega, E.M.M., Cordeiro, G.M. and Hens, N. (2016). A new bimodal flexible distribution for lifetime data. *Journal of Statistical Computation and Simulation*, **88**, 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.
- Silva, M.A., Bezerra-Silva, G.C.D., Vendramim, J.D. and Mastrangelo, T. (2013). Sublethal effect of neem extract on Mediterranean fruit fly adults. *Revista Brasileira de Fruticultura*, **35**, 93–101.
- Silva, G.O., Ortega, E.M.M., Cancho, V.G. and Barreto, M.L. (2008). Log-Burr XII regression models with censored data. *Computational Statistics and Data Analysis*, **52**, 3820–3842.

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.

Therneau, T.M., Grambsch, P.M and Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika*, **77**, 147–160.

Zhu, H., Ibrahim, J.G., Lee, S. and Zhang, H. (2007). Perturbation selection and influence measures in local influence analysis. *The Annals of Statistics*, **35**, 2565–2588.

4 A FLEXIBLE BIMODAL MODEL WITH LONG-TERM SURVIVORS AND DIFFERENT REGRESSION STRUCTURES

Abstract: The cure fraction models are useful to model lifetime time data with long-term survivors. In this paper, we propose a flexible four-parameter cure rate survival model called the exponentiated sinh Cauchy cure rate distribution. We introduce this new distribution in the generalized additive models for location, scale and shape, in order to model any or all the parameters of the distribution using explanatory variables in different regression structures. The maximum likelihood method is used to estimate the model parameters. In order to examine the performance of the proposed model, some simulation are presented to verify the robust aspects of this flexible class against outlying and influential observations. Furthermore, some diagnostic measures and the one-step approximations of the estimates in the case-deletion model are obtained. The flexibility of the proposed model is illustrated by means of three real data sets.

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

4.1 Introduction

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

Models to accommodate a cured fraction have been widely developed. Perhaps the most popular type of cure rate models are the mixture models (MMs) pioneered by Boag (1949), Berkson and Gage (1952) and further studied by Farewell (1982). Recently, the MMs allow both the cure fraction and the survival function of uncured patients (latency distribution) to depend on covariates. Rodrigues *et al.* (2009) developed the COM-Poisson cure rate model considering that the number of competing causes of the event of interest follows the Conway-Maxwell Poisson distribution. Ortega *et al.* (2009) defined the generalized log-gamma regression models with cure fraction to explain/predict the cancer recurrence times. Cancho *et al.* (2013) proposed a destructive negative binomial cure rate model, where the initial number of competing causes of the event of interest follows a compound negative binomial distribution and Hashimoto *et al.* (2014) introduced the Poisson Birbaum-Saunders model with long-term survivors assuming that the number of competing causes of the event of interest follows the Poisson distribution and the time to event has the Birnbaum-Saunders distribution. Recently, Rodrigues *et al.* (2015) studied the relaxed Poisson cure rate model showing an application to cutaneous melanoma data, Ortega *et al.* (2015) proposed a new cure rate survival regression model for predicting breast carcinoma survival in women who underwent mastectomy, Balakrishnan and Pal (2015) derived an EM algorithm for estimation of the parameters of a flexible cure rate model with generalized gamma lifetime and model discrimination using likelihood and information based methods and Balakrishnan *et al.* (2016) proposed piecewise linear approximations for cure rate models and associated inferential issues. Although the models studied in

these papers are attractive, they have some limitations. Most of the proposed models are not able to capture the presence of bi-modality. Another disadvantage is that this model only has a regression structure in the cure fraction.

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 $i = 1, \dots, n$) be the indicator denoting that the i th individual is susceptible ($N_i = 1$) or non-susceptible ($N_i = 0$), i.e., the population is classified in two sub-populations so that an individual either is cured with probability $0 < p < 1$, or has a proper survival function $S(t)$ with probability $(1 - p)$. The mixture model (MM) can be expressed by

$$S_{pop}(t_i) = p + (1 - p)S(t_i|N_i = 1), \quad (4.1)$$

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 $p = P(N_i = 0)$ is the probability of cure of an individual. The probability density function (pdf) corresponding to (4.1) is given by

$$f_{pop}(t_i) = -\frac{dS_{pop}(t_i)}{dt} = (1 - p)f(t_i|N_i = 1), \quad (4.2)$$

where $f(t_i|N_i = 1)$ is the baseline pdf for the susceptible individuals. Equations (4.1) and (4.2) are improper functions, since $S_{pop}(t)$ is not a proper survival function. We can omit sometimes 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 *exponentiated log-sinh Cauchy* (ELSC) distribution, which accommodates various shapes of the skewness, kurtosis and bi-modality. The ELSC density function can be written as

$$f(t; \mu, \sigma, \nu, \tau) = \frac{\tau\nu}{t\sigma\pi} \frac{\cosh\left(\frac{\log(t)-\mu}{\sigma}\right)}{\nu^2 \sinh^2\left(\frac{\log(t)-\mu}{\sigma}\right) + 1} \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh\left(\frac{\log(t)-\mu}{\sigma}\right) \right] \right\}^{\tau-1}, \quad (4.3)$$

where $\mu \in R$ and $\sigma > 0$ are the location and scale parameters, respectively, $\nu > 0$ is the symmetry parameter, which characterizes the bi-modality of the distribution, and $\tau > 0$ is the skewness parameter. The advantage of the ELSC 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 survival function corresponding to (4.3) is given by

$$S(t; \mu, \sigma, \nu, \tau) = 1 - \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh\left(\frac{\log(t)-\mu}{\sigma}\right) \right] \right\}^{\tau}. \quad (4.4)$$

Considering that the failure times follow the ELSC distribution, we propose a new model called the *exponentiated log-sinh Cauchy cure rate* (ELSCcr) model. The paper is organized as follows. In Section 4.2, we propose the ELSCcr model by defining the density and survival functions and discuss inferential issues. We adopt the abbreviation GAMLSS for *generalized additive model for location, scale and shape*. In Section 4.3, we propose the ELSCcr GAMLSS. We also discuss inferential issues, related models and model selection strategies. Some strategies to select the best model, residual analysis, goodness of fit and global influence measure are addressed in Section 4.4. Section 4.5 contains methods for generating random values and two Monte Carlo simulations on the finite sample behavior of the maximum likelihood estimates (MLEs). Applications to three real data sets are presented in Section 4.6 to illustrate the new regression model. Finally, we offer some conclusions in Section 4.7.

4.2 The ELSC model for survival data with long-term survivors

For censored survival times, the presence of an immune proportion of individuals who are not subject to death, failure or relapse may be indicated by a relatively high number of individuals with

large censored survival times. Now, we define the ELSCcr model for the possible presence of long-term survivors in the data. To formulate the model, we consider that the population under study is a mixture of susceptible (uncured) individuals, who may experience the event of interest, and non-susceptible (cured) individuals, who will not experience it (Maller and Zhou, 1996).

4.2.1 Definition

The survival function of the ELSCcr model is defined by assuming that the survival function for susceptible individuals in (4.1) is given by (4.4), which leads to

$$S_{pop}(t; \mu, \sigma, \nu, \tau, p) = 1 + (p - 1) \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan [\nu \sinh(w)] \right\}^{\tau}, \quad (4.5)$$

where $w = \frac{\log(t) - \mu}{\sigma}$. We can omit sometimes the dependence on the parameters to simply as for example, $S_{pop}(t) = S_{pop}(t; \mu, \sigma, \nu, \tau, p)$. The pdf corresponding to (4.5) is given by

$$f_{pop}(t) = \frac{(1-p)\tau\nu}{t\sigma\pi} \frac{\cosh(w)}{[\nu^2 \sinh^2(w) + 1]} \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan [\nu \sinh(w)] \right\}^{\tau-1}. \quad (4.6)$$

The hazard rate function (hrf) of the ELSCcr model is given by $h_{pop}(t) = f_{pop}(t)/S_{pop}(t)$. A random variable having density (4.6) is denoted by $T \sim ELSCcr(\mu, \sigma, \nu, \tau, p)$. 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 ELSCcr survival and hazard functions for selected parameter values are displayed in Figures 4.1 and 4.2, respectively.

Figure 4.1. The ELSCcr survival function when $\mu = 4$, $\sigma = 0.1$ and: (a) For $p = 0$ and different values of ν and τ ; (b) For $p = 0.2$ and different values of ν and τ .

Figure 4.1(a)-(b) reveals clearly the bi-modality and symmetric effects caused by the parameters ν and τ , respectively, and different effects of the cured probability p . Further, Figure 4.2(a) indicates that the hrf of T has decreasing, unimodal, bimodal and unimodal and bathtub-shaped forms. We can note in Figure 4.2(b) that the values of the hrf are smaller in the presence of the proportion of cured but still assuming bimodal characteristics.

4.3 Regression model

In many applications of long term survival models, the cure rate plays an essential role that can be explained by explanatory variables. For example, in medical problems, the lifetimes and the cure rate are affected by the cholesterol level, blood pressure, weight and many others. Parametric models to estimate univariate survival functions for censored data regression problems are widely used. Recently,

Figure 4.2. The ELSCr hrf for different values of μ , σ , ν and τ and: (a) $p = 0$ and (b) $p = 0.2$.

several regression models for long-term survivors have been proposed in the literature, as mentioned in Section 1. In general, these models assume only that the cure rate “ p ” and location “ μ ” parameters must be modeled by explanatory variables. A disadvantage of the class of location models is that the variance, skewness, bi-modality, kurtosis and other parameters are not modelled explicitly in terms of the explanatory variables. As an alternative, the systematic part of the GAMLSS (Rigby and Stasinopoulou, 2005) can be expanded to allow not only the location but all the parameters of the conditional distribution of T to be modelled as parametric functions of the explanatory variables.

4.3.1 Parametric model

Let $T \sim \text{ELSCr}(y; \theta)$, where $\theta^T = (\mu, \sigma, \nu, \tau, p)$ denotes the vector of parameters of the pdf (4.6). Consider independent observations t_i conditional on the parameter vector θ_i (for $i = 1, 2, \dots, n$) having pdf $f(t_i; \theta_i)$, where $\theta^T = (\mu^T, \sigma^T, \nu^T, \tau^T, \mathbf{p}^T)$ is a vector of parameters related to the response variable. We can define the elements of the vector θ using appropriate link functions as

$$\begin{aligned} g_1(\mu) &= \mathbf{X}_1 \beta_1, & g_2(\sigma) &= \mathbf{X}_2 \beta_2, \\ g_3(\nu) &= \mathbf{X}_3 \beta_3, & g_4(\tau) &= \mathbf{X}_4 \beta_4, & g_5(\mathbf{p}) &= \mathbf{X}_5 \beta_5, \end{aligned} \quad (4.7)$$

where $g_k(\cdot)$ for $k = 1, 2, 3, 4, 5$, denote the injective and twice continuously differentiable monotonic link functions, $\beta_k = (\beta_{0k}, \beta_{1k}, \dots, \beta_{m_k k})^T$ is a parameter vector of length $(m_k + 1)$, m_k denotes the number of explanatory variables related to the k th 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 defined by $m = m_1 + m_2 + m_3 + m_4 + m_5 + 5$ and the choice of parameters to be modeled by explanatory variables is discussed in Section 4.3.4. For the following sections, we shall consider the identity link function for $g_1(\cdot)$ and the logarithmic link function for $g_k(\cdot)$ ($k = 2, 3, 4$).

We emphasize that estimating the proportion of cure is very important since most researchers adopt the logit link for the structure regression of the cure fraction. In this study, in addition using the logit link, which is usual in long-term survivors, we propose the logit, complementary log-log, log-log and probit links for $g_5(\cdot)$, as specified below:

- Logit link: $\mathbf{p} = \frac{\exp(\mathbf{X}_5 \beta_5)}{[1 + \exp(\mathbf{X}_5 \beta_5)]}.$
- Complementary log-log link: $\mathbf{p} = 1 - \exp[-\exp(\mathbf{X}_5 \beta_5)].$
- Log-log link: $\mathbf{p} = \exp[-\exp(-\mathbf{X}_5 \beta_5)].$
- Probit link: $\mathbf{p} = \Phi(\mathbf{X}_5 \beta_5),$

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

4.3.2 Related models

Let $T \sim ELSCr(\mu, \sigma, \nu, \tau, p)$ be a random variable having density function (4.6). Sub-models and related distributions are listed in Table 4.1. Note that for $p \neq 0$ all models described above are extended to models with cure rate, e.g., for $p \neq 0$, $\sigma = 1$, $\tau = 1$ and μ , we obtain the *folded Cauchy cure rate* (FCcr) model.

Table 4.1. Related distributions.

Distribution	μ	σ	ν	τ	p	References
Exponentiated log-sinh Cauchy (ELSC)	μ	σ	ν	τ	0	Ramires <i>et al.</i> (2016)
Log-sinh Cauchy (LSC)	μ	σ	ν	1	0	Ramires <i>et al.</i> (2016)
Folded Cauchy (FC)	$\log(\mu)$	1	ν	1	0	Johnson <i>et al.</i> (1994)
for $X = \log(T)$						
Exponentiated sinh Cauchy (ESC)	μ	σ	ν	τ	0	Cooray (2013)
Sinh Cauchy (SC)	μ	σ	ν	1	0	Cooray (2013)
Hyperbolic secant (HS)	μ	σ	1	1	0	Talacko (1956)

Further, we work with the parametric regression model (4.7). This stricture can be used to extend all sub-models presented above in the GAMLSS class, e.g., for $\mathbf{p} = \mathbf{0}$, we obtain the ELSC GAMLSS regression. We can note that the GAMLSS family extends two important and usual classes of regression models. The class of location models is obtained by taking $m_2 = m_3 = m_4 = 0$ and, for $m_3 = m_4 = 0$, $m_1 \neq 0$ and $m_2 \neq 0$, we have the regression model with heteroscedastic errors.

4.3.3 Inference

Consider a sample of n -independent observations $t_1, \dots, t_n$. Let c_i denote the censoring time, $t_i = \min\{t_i, c_i\}$ 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 $(t_1, \delta_1, \mathbf{x}_{k1}), \dots, (t_n, \delta_n, \mathbf{x}_{kn})$, the log-likelihood function under non-informative censoring is given by

$$\begin{aligned}
 l(\boldsymbol{\theta}) &= \sum_{i \in F} \left\{ \log(1 - p_i) + \log(\tau_i \nu_i) - \log(\sigma_i \pi) - \log(t_i) + \log \cosh(w_i) - \log[1 + \nu_i^2 \sinh^2(w_i)] \right\} \\
 &\quad + \sum_{i \in F} (\tau_i - 1) \log \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu_i \sinh(w_i)] \right\} \\
 &\quad + \sum_{i \in C} \log \left(1 + (p_i - 1) \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu_i \sinh(w_i)] \right\}^{\tau_i} \right), \tag{4.8}
 \end{aligned}$$

where the parameter vector $\boldsymbol{\theta} = (\beta_1^T, \dots, \beta_5^T)^T$, and the parameters β_k , $k = 1, \dots, 5$ are defined in (4.7) by specifying appropriate link functions for $g_k(\cdot)$, e.g, using the logit link function for $g_5(\mathbf{p})$, the parameter p is related to the covariates by replacing p by $p_i = \exp(\mathbf{X}_5[i,] \beta_5) / [1 + \exp(\mathbf{X}_5[i,] \beta_5)]$, where $\mathbf{X}_k[i,]$ denotes the i -th row of the model matrix $\mathbf{X}_k$. Then, the score functions for the parameters in $\boldsymbol{\theta}$ are given by

$$\begin{aligned}
 \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_{j1}} &= \sum_{i \in F} \left[\dot{g}_1^{-1}(\mu_i) \right]_{\beta_{j1}} \left[\frac{\nu_i^2 \sinh(2w_i)}{\sigma_i K_i} - \frac{\tanh(w_i)}{\sigma_i} - (\tau_i - 1) \frac{\nu_i \cosh(w_i)}{\pi \sigma J_i K_i} \right] \\
 &\quad - \sum_{i \in C} \left[\dot{g}_1^{-1}(\mu_i) \right]_{\beta_{j1}} \frac{(p_i - 1) \tau_i \nu_i \cosh(w_i) J_i^{\tau_i - 1}}{\pi_i \sigma_i K_i (1 + (p_i - 1) J_i^{\tau_i})}, \\
 \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_{j2}} &= \sum_{i \in F} \left[\dot{g}_2^{-1}(\sigma_i) \right]_{\beta_{j2}} \left[\frac{1 - w_i \tanh(w_i)}{\sigma} + \frac{\nu^2 w_i}{\sigma K_i} \sinh(2w_i) - (\tau - 1) \frac{\nu w_i \cosh(w_i)}{\pi \sigma J_i K_i} \right] \\
 &\quad - \sum_{i \in C} \left[\dot{g}_2^{-1}(\sigma_i) \right]_{\beta_{j2}} \frac{\tau \nu w_i J_i^{\tau - 1} \cosh(w_i)}{\pi \sigma K_i (1 + (p - 1) J_i^{\tau})}, \\
 \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_{j3}} &= \sum_{i \in F} \left[\dot{g}_3^{-1}(\nu_i) \right]_{\beta_{j3}} \left[\frac{1}{\nu_i} - \frac{2\nu_i \sinh^2(w_i)}{K_i} + (\tau_i - 1) \frac{\sinh(w_i)}{\pi J_i K_i} \right] \\
 &\quad + \sum_{i \in C} \left[\dot{g}_3^{-1}(\nu_i) \right]_{\beta_{j3}} \frac{(p_i - 1) \tau_i J_i^{\tau_i - 1} \sinh(w_i)}{\pi K_i (1 + (p_i - 1) J_i^{\tau_i})},
 \end{aligned}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_{j_4^4}} = \sum_{i \in F} [\dot{g}_4^{-1}(\tau_i)]_{\beta_{j_4^4}} \left[\frac{1}{\tau_i} + \log(J_i) \right] + \sum_{i \in C} [\dot{g}_4^{-1}(\tau_i)]_{\beta_{j_4^4}} \frac{(p_i - 1)J_i^{\tau_i}}{1 + (p_i - 1)J_i^{\tau_i}} \log(J_i) \quad \text{and}$$

$$\frac{\partial l(\boldsymbol{\theta}_i)}{\partial \beta_{j_5^5}} = \sum_{i \in F} [\dot{g}_5^{-1}(p_i)]_{\beta_{j_4^4}} \frac{(-1)}{1 - p_i} + \sum_{i \in C} [\dot{g}_5^{-1}(p_i)]_{\beta_{j_4^4}} \frac{J_i^{\tau_i}}{1 + (p_i - 1)J_i^{\tau_i}},$$

where $[\dot{g}_k^{-1}(\cdot)]_{\beta_{j_k k}} = \frac{\partial [g_k^{-1}(\cdot)]}{\partial \beta_{j_k k}}$, for $k = 1, \dots, 5$ and $j_k = 0, 1, \dots, m_k$, $J_i = \frac{1}{2} + \frac{1}{\pi} \arctan[\nu_i \sinh(w_i)]$, $K_i = \nu_i^2 \sinh^2(w_i) + 1$ and $w_i = [\log(y_i) - \mu_i]/\sigma_i$.

The numerical maximization of the log-likelihood can be performed in the R software, and the `manipulate` package can be used to define the initial parameter values.

The fit of the ELSCr model gives the estimated survival function

$$\hat{S}_{pop}(t_i) = 1 + (\hat{p}_i - 1) \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan [\hat{\nu}_i \sinh(\hat{w}_i)] \right\}^{\hat{\tau}_i},$$

where

$$\hat{w}_i = \frac{\log(t_i) - \hat{\mu}_i}{\hat{\sigma}_i}, \quad \hat{\mu}_i = g_1(\mathbf{x}_{1i}^T \hat{\boldsymbol{\beta}}_1), \quad \hat{\sigma}_i = g_2(\mathbf{x}_{2i}^T \hat{\boldsymbol{\beta}}_2), \quad \hat{\nu}_i = g_3(\mathbf{x}_{3i}^T \hat{\boldsymbol{\beta}}_3), \quad \tau_i = g_4(\mathbf{x}_{4i}^T \hat{\boldsymbol{\beta}}_4),$$

$$p_i = g_5(\mathbf{x}_{5i}^T \hat{\boldsymbol{\beta}}_5), \quad \hat{\boldsymbol{\beta}}_k = (\hat{\beta}_{0k}, \hat{\beta}_{1k}, \dots, \hat{\beta}_{m_k k})^T, \quad \mathbf{x}_{ki}^T = (1, x_{1ik}, \dots, x_{im_k k})$$

for $k = 1, \dots, 5$, $i = 1, \dots, n$ and $g_k(\cdot)$ is defined in Section 3.

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

4.3.4 Selecting explanatory variables and link functions

For the ELSCr GAMLSS regression, the selection of the terms for all parameters is performed using a stepwise of the generalized Akaike information criterion (GAIC) procedure (see Section 4.4.1 for details of the GAIC). There are many different strategies that could be applied for selection of the terms used to model the five parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$, $\boldsymbol{\tau}$ and $\mathbf{p}$. Here, we adopt a modification of the strategy described by Voudouris *et al.* (2012). Let χ be the selection of all terms available for consideration, where χ contains the linear terms. Then, for all terms in χ and for fixed distribution and link functions, the strategy consists in two steps. In the first step, we adopt a forward selection procedure to select an appropriate model for $\boldsymbol{\mu}$, with $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$, $\boldsymbol{\tau}$ and $\mathbf{p}$ fitted as constants. After then, repeat the same procedure to select the model for $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$, $\boldsymbol{\tau}$ and $\mathbf{p}$, respectively, using the models already obtained in the previous steps as constants. For the second step, we perform a backward selection procedure to choose an appropriate model for $\boldsymbol{\tau}$, with $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\mathbf{p}$ fitted as constants and repeat this procedure for $\boldsymbol{\nu}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$, respectively. At the end of the steps described above, the final model may contain different subsets from χ for $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$. On the other hand, the choice of the link functions can be done using the GAIC statistic or can also be fixed to facilitate interpretation of the parameters.

4.4 Goodness of fit, diagnostics and influence measures

There exist a variety of methodologies to compare several competing models for a given data set and select the one that provides the best fit to the data. The selection of the appropriate distribution is performed in two stages, the fitting stage and the diagnostic stage. In the first stage, the GAIC measure is used to compare different fitted models. The model with the smallest value of the $\text{GAIC}(k)$ criterion is selected. The diagnostic stage involves the use of residual plots to study departures from the error

assumption and the presence of outliers. In the diagnostic stage, we can also use influence measures to find those models most affected by atypical observations. These two stages can be adopted for all models presented in Sections 4.2 and 4.3.

4.4.1 Choosing the best model

The GAIC is defined by $\text{GAIC}(k) = GD + k \times df$, where GD represents the global deviance $GD = -2l(\hat{\theta})$, $l(\hat{\theta})$ is the maximized log-likelihood function, df is the total effective degrees of freedom of the fitted model and k is a constant. The model with the smallest value of the $\text{GAIC}(k)$ criterion is then selected. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are special cases of the $\text{GAIC}(k)$ measure corresponding to $k = 2$ and $k = \log(n)$, respectively. The AIC and BIC statistics are asymptotically justified for predicting the goodness-of-fit to the current data, that is, approximations to the average predictive error. We opted to the AIC and BIC criteria to select the best models. We can also use the likelihood ratio (LR) statistic for comparing some nested models. The LR statistic for testing the hypotheses $H_0 : \theta = \theta_1$ versus $H_1 : \theta \neq \theta_1$ is given by $w = 2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_1)\}$, where θ_1 is a specified vector and $\hat{\theta}$ and $\hat{\theta}_1$ are the estimates under the null and alternative hypotheses, respectively. The statistic w is asymptotically (as $n \rightarrow \infty$) distributed as χ_q^2 , where q is the difference in dimensionality of θ and θ_1 .

4.4.2 Diagnostic and influence analysis

Since regression models are sensitive to the underlying model assumptions, generally performing a sensitivity analysis is strongly advisable. In order to study departures from the error assumption and the presence of outliers, we can use the normalized randomized quantile residuals (Dunn and Smyth, 1996). These residuals can be easily determined by $\hat{r}_i = \Phi^{-1}(\hat{u}_i)$, where $\Phi^{-1}(\cdot)$ is the inverse cdf of the standard normal variate, $\hat{u}_i = 1 - S(y_i|\hat{\theta}_i)$ and $S(y_i|\hat{\theta}_i)$ is the survival function (4.5). For censored response variables, considering a right censored continuous response, $\hat{u}$ is defined as a random value from a uniform distribution on the interval $[1 - S(y_i|\hat{\theta}_i), 1]$.

Since regression models are sensitive to the underlying model assumptions, performing a sensitivity analysis is strongly advisable. Cook (1986) used this idea to motivate the assessment of influence analysis. The best known perturbation schemes are based on case-deletion (Cook and Weisberg, 1982), 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 (4.7), the log-likelihood function of θ is denoted by $l_{(-i)}(\theta)$. Let $\hat{\theta}_{(-i)}^T = (\hat{\mu}_{(-i)}^T, \hat{\sigma}_{(-i)}^T, \hat{\nu}_{(-i)}^T, \hat{\tau}_{(-i)}^T, \hat{\mathbf{p}}_{(-i)}^T)$ be the MLEs of the parameters from $l_{(-i)}(\theta)$. To assess the influence of the i th case on the MLEs, the basic idea is to compare the difference between $\hat{\theta}_{(-i)}$ and $\hat{\theta}$. If deletion of a case seriously influences the estimates, for example, changing the inference, more attention should be given to that case. Hence, if $\hat{\theta}_{(-i)}$ is far from $\hat{\theta}$, then the i th case is regarded as an influential observation. A popular measure of the difference between $\hat{\theta}_{(-i)}$ and $\hat{\theta}$ is the log-likelihood distance defined by

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

where $l(\hat{\theta})$ is given by (4.8). For a specific data set and model, the penalized log-likelihood can potentially have multiple local maxima, so we suggest the MLE $\hat{\theta}^T$ as initial trial vector to obtain the estimate $\hat{\theta}_{(-i)}^T$.

4.5 Simulation

We simulate ELSCcr random variables by inverting $F(t) = 1 - S(t) = u$ in (4.4). We obtain the quantile function (qf) of $T \sim \text{ELSC}(\mu, \sigma, \nu, \tau)$ by

$$T = Q(u) = \exp \left(\mu + \sigma \operatorname{arcsinh} \left\{ \frac{1}{\nu} \tan \left[\pi \left(u^{1/\tau} - 0.5 \right) \right] \right\} \right). \quad (4.9)$$

Equation (4.9) 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. The cured proportion can be generated using the qf of another distribution with real support, fixing p and setting the sample size for cured individuals as $n_c = p \times n$. We can also simulate the regression models setting the parameters using the parametric structure (4.7).

Here, we conduct two Monte Carlo simulation studies to assess the finite sample behavior of the MLEs of the parameters for different sample sizes and cure rate percentages. In the first simulation, we consider the model presented in Section 4.2 and, in the second simulation, we consider the GAMLSS regression (4.7) by modeling all parameters using the explanatory variables. In the two simulation studies, the sample sizes are generated by taking $n = 50$ and 100 , where the failure times T are generated from the ELSC distribution using the qf (4.9) and the censoring times, denoted by C , are randomly generated from the uniform distribution $C \sim U(200, 250)$.

The lifetimes considered in each fit are evaluated as $\min(t_i, c_i)$ and for each configuration of n and p , 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 determine the average estimates (AEs), biases and means squared errors (MSEs). The simulations are carried out using the R programming language, where the `optim` algorithm is used for maximizing the total log-likelihood function (4.8).

4.5.1 Simulation 1: ELSCcr model

We simulate the ESCcr distribution (for $\mu = 4, \sigma = 0.2, \nu = 0.1, \tau = 1$ and $p = 0, 0.3, 0.5$), considering bi-modality form. The results of the Monte Carlo study are given in Table 4.2. They indicate that the MSEs of the MLEs of the parameters decay toward zero as the sample size increases, as expected under first-order asymptotic theory.

Table 4.2. The AEs, biases and MSEs based on 1,000 simulations for the ESCcr model when $\mu=4$, $\sigma = 0.2$, $\nu = 0.1$, $\tau = 2$ and $p = 0, 0.3, 0.5$, for $n=50$ and 100 .

p	Parameter	$n = 50$			Parameter	$n = 100$		
		AE	Bias	MSE		AE	Bias	MSE
0.0	μ	4.008	0.008	0.002	μ	4.008	0.008	0.001
	σ	0.167	-0.033	0.002	σ	0.169	-0.031	0.001
	ν	0.069	-0.031	0.002	ν	0.070	-0.030	0.002
	τ	1.274	0.274	0.124	τ	1.268	0.268	0.095
	p	0.000	0.000	0.001	p	0.000	0.000	0.001
0.3	μ	4.013	0.013	0.004	μ	4.010	0.010	0.002
	σ	0.176	-0.024	0.002	σ	0.178	-0.022	0.001
	ν	0.081	-0.019	0.003	ν	0.080	-0.020	0.002
	τ	1.324	0.324	0.178	τ	1.307	0.307	0.130
	p	0.283	-0.017	0.001	p	0.285	-0.015	0.001
0.5	μ	4.016	0.016	0.007	μ	4.009	0.009	0.002
	σ	0.176	-0.024	0.002	σ	0.177	-0.023	0.001
	ν	0.085	-0.015	0.006	ν	0.080	-0.020	0.002
	τ	1.351	0.351	0.250	τ	1.316	0.316	0.149
	p	0.485	-0.015	0.002	p	0.490	-0.010	0.001

4.5.2 Simulation 2: ELSCcr regression model

For the ELSCcr GAMLSS regression, we consider the lifetimes T composed by the lifetimes of two groups T_1 and T_2 , where the groups g_1 and g_2 are represented by the explanatory variable $x_{1i} = 1$ and $x_{1i} = 0$, respectively, for $i = 1, \dots, n$. Consider different characteristics for each group, such as location, scale, asymmetry, bi-modality and cured proportion. We define

$$T_1 \sim \text{ELSCcr}(4.5, 0.135, 0.606, 1.221, 0.269) \quad \text{and} \quad T_2 \sim \text{ELSCcr}(4, 0.082, 0.011, 1, 0.119),$$

and with this configuration, the true parameter values used in the data-generating processes are:

$$\begin{aligned} \mu_i &= 4 + 0.5x_{1i}, & \sigma_i &= \exp(-2.5 + 0.5x_{1i}), & \nu_i &= \exp(-4.5 + 4x_{1i}), \\ \tau_i &= \exp(0 + 0.2x_{1i}) & \text{and} & & p_i &= \frac{\exp(-2 + x_{1i})}{\exp(-2 + x_{1i}) + 1}. \end{aligned}$$

The results are reported in Table 4.3 and, for visual analysis, we present in Figure 4.3 the generated and the estimated (considering the AEs given in Table 4.3) survival functions for $n = 50$ and 100 and considering the two groups represented by the variable x_i .

Table 4.3. The AEs, biases and MSEs based on 1,000 simulations of the ELSCcr GAMLSS regression when $\beta_{01}=0.5$, $\beta_{11}=6$, $\beta_{02} = 1.5$, $\beta_{12} = 0.6$, $\beta_{03} = -3.5$, $\beta_{13} = 3$, $\beta_{04} = 0.2$ and $\beta_{14} = 0.9$, for $n = 50$ and 100 and under censoring percentages $\kappa = 0.0, 0.1$ and 0.3 .

Parameter	$n = 50$			Parameter	$n = 100$		
	AE	Bias	MSE		AE	Bias	MSE
β_{01}	3.999	-0.001	0.001	β_{01}	3.999	-0.001	0.001
β_{11}	0.443	-0.057	0.038	β_{11}	0.454	-0.046	0.032
β_{02}	-2.567	-0.067	0.038	β_{02}	-2.535	-0.035	0.018
β_{12}	0.465	-0.035	0.505	β_{12}	0.502	0.002	0.185
β_{03}	-4.928	-0.428	1.388	β_{03}	-4.733	-0.233	0.590
β_{13}	4.202	0.202	2.657	β_{13}	4.113	0.113	0.959
β_{04}	0.002	0.002	0.074	β_{04}	-0.002	-0.002	0.037
β_{14}	0.496	0.296	1.418	β_{14}	0.398	0.198	0.723
β_{05}	-2.112	-0.112	0.032	β_{05}	-1.971	0.029	0.008
β_{15}	1.105	0.105	0.345	β_{15}	0.994	-0.006	0.038

Figure 4.3. Some ELSCcr survival functions at the true parameter values and at the AEs obtained in Table 4.3 by taking (a) $n = 50$ and (b) $n = 100$.

The results of the Monte Carlo study in Tables 4.2 and 4.3 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 oftentimes be improved by using bias adjustments to these estimators. In general, for the ESC regression models, the variances and MSEs increase when the censoring percentage increases. This fact can be noted in Figure 4.3.

4.6 Applications

In this section, we provide three applications to real data to prove empirically the flexibility of the ELSCcr model. In the first application, we show the flexibility of the ELSCcr distribution defined in Section 4.2. The second and third applications prove empirically the usefulness of the ELSCcr GAMLSS regression by modeling all/some parameters with explanatory variables. For the three examples presented in this section, the computations are performed using the `optim` subroutine in the R software and the computational codes can be downloaded from <https://goo.gl/5Cd8Ug>.

4.6.1 Calving data

For the first example, we consider the data relative to the ages of the cows at first calving. This data were obtained from the zootechnics records of a Brazilian company engaged in raising beef cattle, located in the states of Bahia and São Paulo. The age at first calving is the main characteristic analyzed, which is an important characteristic for beef cattle breeders due to the fact the faster cows reach reproductive maturity and generating fast return on investment. In this case, the response variable t_i is the age of the cows at first calving (measured in days).

The sample size in this study is $n = 1,326$, where 32.35% of the observations do not present the event of interest (calving) and are thus censored. It is known that time to first calving can be influenced by variables, but for this example, we will only consider the response and censored times. First, we consider that the response variable t_i follows the ELSCcr (4.6) distribution. Then, we compare the results by fitting the LSCcr model, a special case of a ELSCcr model when $\tau = 1$. Table 4.4 lists the MLEs and their corresponding standard errors (SEs) in parentheses of the model parameters and the values of the AIC and BIC statistics for the fitted models.

Table 4.4. MLEs of the model parameters for the calving data, the corresponding SEs (given in parentheses) and the AIC and BIC statistics.

Model	μ	σ	ν	τ	p	AIC	BIC
ELSCcr	6.844 (0.002)	0.029 (0.001)	0.023 (0.003)	1.637 (0.069)	0.323 (0.013)	11955.9	11981.9
LSCcr	6.855 (0.001)	0.026 (0.001)	0.019 (0.002)	1 -	0.320 (0.012)	12080.2	12101.0

The figures in Table 4.4 indicate that the ELSCcr model has the lowest AIC and BIC values, and therefore it could be chosen as the best model. Further, using the LR statistic to compare the fits of these models, i.e., for testing the null hypothesis $H_0 : \tau = 1$, we obtain $w = 126.33$ with the p -value < 0.001 . Then, we could accept the ELSCcr distribution. In order to verify the adequacy and the assumptions of the ELSCcr model, the index plot for the quantile residuals is displayed in Figure 4.4. We note in this plot eight points (0.6% of the sample) out of the range $[-3, 3]$, which represent the eight smaller ages at first calving.

The adequacy of the fitted ELSCcr and LSCcr models can be noted in Figure 4.5(a), which gives the empirical and estimated survival functions for the current data. We also present in Figure 4.5(b) the fitted hazard function for the ELSCcr model, where the presence of the bi-modality is evident. In general, we can conclude that the ELSCcr distribution provides a good fit to these data.

Figure 4.4. For calving data, the index plot of quantile residuals.

Figure 4.5. For calving data, (a) the estimated and empirical survival function for the ELSCcr and LSCcr models and (b) the estimated hazard function for the ELSCcr model.

4.6.2 Gastric cancer data

Gastric cancer is one of the leading causes of cancer-related death and the mucosal resection is accepted as a treatment option for early cases of the disease. It is known that the chemoradiotherapy (CRT) is the standard treatment used for gastric cancer patients. On the other hand, new technologies to optimize medical decisions and the development of new therapies are of great importance to improve survival in gastric cancer. Therefore, J come *et al.* (2013) conducted a study in patients with gastric adenocarcinoma who underwent curative resection in which was compared the 3 year overall survival of the two treatments. The study consisted of $n = 201$ patients of different clinical stages, which includes 76 patients that received adjuvant CRT and 125 that received resection alone. Here, the response variable T refers to the lifetimes in months since surgery and the treatments resection alone and CRT is represented by $X_1 = 0$ and $X_1 = 1$, respectively. We consider censored the lifetimes of the patients who remain alive after the end of the study. These data are obtained in Martinez *et al.* (2013).

We start the analysis by fitting the ELSCcr regression model (4.7). Using the steps described in Section 4.3.4 to select the additive terms for the different parameters and considering different link functions for the $\mathbf{p}$ parameter factor, we present results for the model parameters defined by

$$\mu_i = \beta_{01}, \sigma_i = \exp(\beta_{02}), \nu_i = \exp(\beta_{03} + \beta_{13}x_{i1}), \tau_i = \exp(\beta_{04} + \beta_{14}x_{i1}) \text{ and } g_5(p_i) = \beta_{05} + \beta_{15}x_{i1},$$

where $g_5(\cdot)$ can be taken as the logit, complementary log-log, log-log or probit link functions. Table 4.5 lists the values of the AIC and BIC statistics for the fitted models under different link functions. We conclude that the log-log link function gives the lowest values of AIC and BIC statistics. Table 4.6

provides the MLEs, SEs and p -values obtained from the fitted ELSCcr regression model taking log-log link function for $g_5(\cdot)$. We note that the parameter β_{15} is not significant at 5%, indicating that we do not have evidence of differences between the population cure fractions considering patients treated by adjuvant chemoradiotherapy and surgery alone.

Table 4.5. The AIC and BIC statistics for the fitted models to the gastric data under different link functions for $\mathbf{p}$.

Link functions for g_5	AIC	BIC
logit	869.4	895.9
complementary log-log	869.5	896.0
log-log	869.3	895.7
probit	869.7	896.2

Table 4.6. For the Gastric cancer data, the MLEs and the corresponding SEs and p -values of the estimates from the fitted ELSCcr regression model by taking the log-log link function for $g_5(\cdot)$.

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

To verify the adequacy and the assumptions of the fitted model in Table 4.6, we present in Figure 4.6(a) the index plots for the quantile residuals. We also present in Figure 4.6(b) the case deletion measure $LD_i(\boldsymbol{\theta})$. We may observe in these plots that the quantile residuals follow approximately a normal distribution and has not been identified a possibly influential observation.

Figure 4.6. For gastric cancer data, the index plot of (a) quantile residuals and (b) the absolute values of likelihood distance.

In order to assess if the model is appropriate, the empirical and estimated survival function of the ELSCcr regression model are plotted in Figure 4.7(a). We also present in Figure 4.7(b) the estimated hazard functions, which reveal that the hazard of death is higher in the time immediately after the surgery considering the patients that received the surgery alone. In other hand, for the patients that received the chemoradiotherapy, the hazard of death has bimodal form with high values at 15 and 27 months after the surgery intervention. We conclude that the ELSCcr regression model provides a good fit to these data.

Figure 4.7. For gastric cancer data, (a) the estimated and empirical survival functions and (b) the estimated hazard functions.

4.6.3 Breast cancer data

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

To construct the data set used in this example, we used five data sets that are available as experimental data packages on Bioconductor.org. Molecular information has been extracted from the phenotype (pData) of the corresponding data set under the Gene Expression Omnibus (GEO) and to perform molecular sub-typing, we adopt the SCMOD2 sub-typing algorithms. The steps to construct these data can be found in Gendoo *et al.* (2015).

The final data consist of $n = 493$ observations containing the lifetime t_i (in months) of patients as well the breast cancer subtypes, which are represented by dummies variables as follows: Basal ($X_1 = 0, X_2 = 0, X_3 = 0$), Her2 ($X_1 = 1, X_2 = 0, X_3 = 0$), Lumial A ($X_1 = 0, X_2 = 1, X_3 = 0$) and Lumial B ($X_1 = 0, X_2 = 0, X_3 = 1$). After performing the model selection described in Section 4.3.4 to select the terms of the regression structure (4.7), we present results where the model parameters are defined by

$$\begin{aligned} \mu_i &= \beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2} + \beta_{31}x_{i3}, & \sigma_i &= \exp(\beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2} + \beta_{32}x_{i3}), \\ \nu_i &= \exp(\beta_{03} + \beta_{13}x_{i1} + \beta_{23}x_{i2} + \beta_{33}x_{i3}), & \tau_i &= \exp(\beta_{04} + \beta_{14}x_{i1} + \beta_{24}x_{i2} + \beta_{34}x_{i3}) \quad \text{and} \\ g_5(p_i) &= \beta_{05} + \beta_{15}x_{i1} + \beta_{25}x_{i2} + \beta_{35}x_{i3}, \end{aligned}$$

where $g_5(\cdot)$ can be represented by the logit, complementary log-log, log-log or probit link functions. To select the best link function, we present in Table 4.7 the values of the AIC and BIC statistics for the fitted models under different link functions for $g_5(\cdot)$. We conclude that the logit link function gives the lowest values of the AIC and BIC statistics.

Table 4.8 provides the MLEs, SEs and p -values obtained from the fitted ELSCr regression model. We may note that β_{25} is significant at the 1% level, indicating a difference between the population cure rate fractions of Lumial A and Basal subtypes. We can also note that the subtypes have a significant effect on the location, scale, skewness and bi-modality parameters, so it should be used to obtain accurate estimates.

The index plots for the quantile residuals is displayed in Figure 4.8(a) in order to verify the adequacy and the assumptions of the proposed model. We may note in this plot that the quantile residuals

Table 4.7. The AIC and BIC statistics for the fitted models to the breast data considering different link functions for $\mathbf{p}$.

Link functions for g_5	AIC	BIC
logit	1799.9	1883.9
complementary log-log	1800.0	1884.0
log-log	1803.5	1887.5
probit	1801.1	1885.1

Table 4.8. MLEs of parameters, degree of freedom and the approximate SEs from the fitted semiparametric ESC and normal models to the body mass data.

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

follow approximately a normal distribution and that the observation #447 appears as a possible outlier. On the other hand, Figure 4.8(b) reveals the case deletion measure $LD_i(\boldsymbol{\theta})$ and again the #447 case appears as a possibly influential observation. In fact, it represents the lowest value of lifetimes for the Lumial A subtype.

Figure 4.8. For breast cancer data, the index plots for (a) quantile residuals and (b) the absolute values of likelihood distance.

The adequacy of the fits can also be observed in Figure 4.9, which presents the empirical and estimated survival function for each breast cancer subtypes. The fitted hazard functions are also given in Figure 4.10, where we observe bimodal shapes for the Basal, Her2 and Lumial B subtypes. These plots evidence the non-proportionality of the hazard functions, making attractive the use of parametric models for the analysis of these data since they do not consider the assumption of proportional hazards used in the usual semi-parametric Cox model. We can conclude that the ELSCcr regression model yields a good fit for the breast cancer data.

Figure 4.9. For breast cancer data, the estimated and empirical survival functions for (a) Basal, Lumial A, (b) Her2 and Lumial B subtypes.

Figure 4.10. For breast cancer data, the estimated hazard functions for (a) Basal, Lumial A, (b) Her2 and Lumial B subtypes.

4.7 Conclusions

We propose the exponentiated log-sinh Cauchy cure rate (ELSCcr) model that can be used as an alternative to mixture distributions in modeling bimodal data with or without the presence of immune proportion of individuals. We show that it can accommodate various shapes of the skewness, kurtosis and bi-modality. We also provide regression structures for all parameters related to location, scale, bi-modality and skewness, which are expressed as linear functions of explanatory variables. Some numerical experiments reveal that the maximum likelihood estimation procedure works well. Three real data examples prove empirically that the ELSCcr distribution is very flexible, parsimonious, and a competitive model that deserves to be added to existing distributions in modeling bimodal data.

References

- Balakrishnan, N. and Pal, S. (2015). An EM algorithm for the estimation of flexible cure rate model parameters with generalized gamma lifetime and model discrimination using likelihood- and information-based methods. *Computational Statistics*, **30**, 151–189.
- Balakrishnan, N., Koutras, M.V., Milienos, F. and Pal, S. (2016). Piecewise linear approximations for cure rate models and associated inferential issues. *Methodology and Computing in Applied Probability*. DOI 10.1007/s11009-015-9477-0 (to appear).

- 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.
- Cancho, V.G., Bandyopadhyay, D., Louzada, F. and Yiqi, B. (2013). The destructive negative binomial cure rate model with a latent activation scheme. *Statistical Methodology*, **13**, 48–68.
- Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society*, **48**, 133–169.
- Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hill.
- Cooray, K. (2013). Exponentiated Sinh Cauchy Distribution with Applications. *Communications in Statistics-Theory and Methods*, **42**, 3838–3852.
- Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.
- Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, **38**, 1041–1046.
- Gendoo, D.M.A., Ratanasirigulchai, N., Schröder, M., Pare, L., Parker, J.S., Prat, A. and Haibe-Kains, B. (2015). *genefu*: a package for breast cancer gene expression analysis. Retrieved 2016-03-30, from <https://bioc.ism.ac.jp/packages/devel/bioc/vignettes/genefu/inst/doc/genefu.pdf>
- Haque, R., Ahmed, S.A., Inzhakova, G., Shi, J., Avila, C., Polikoff, J., Bernstein, L., Enger, M.S. and Press, M.F. (2012). Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. *Cancer Epidemiology Biomarkers & Prevention*, **21**, 1848–1855.
- Hashimoto, E.M., Ortega, E.M.M., Cordeiro, G.M. and Cancho, V.G. (2014). The Poisson Birnbaum-Saunders model with long-term survivors. *Statistics*, **48**, 1394–1413.
- Jcome, A.A.A., Wohnrath, D.R., Neto, C.S., Fregnani, J.H.T.G., Quinto, A.L., Oliveira, A.T.T., Vazquez, V.L., Fava, G., Martinez, E.Z. and Santos, J.S. (2013). Effect of adjuvant chemoradiotherapy on overall survival of gastric cancer patients submitted to D2 lymphadenectomy. *Gastric Cancer*, **16**, 233–238.
- Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). Continuous univariate distributions, vol. 1-2, Wiley.
- Maller, R.A. and Zhou, X. (1996). Survival analysis with long-term survivors. New York: Wiley.
- Martinez, E.Z., Achcar, J.A., Jcome, A.A.A. and Santos, J.S. (2013). Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data. *Computer methods and programs in biomedicine*, **112**, 343–355.
- Ortega, E.M.M., Cancho, V.G. and Paula, G.A. (2009). Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis*, **15**, 79–106.
- Ortega, E.M.M., Cordeiro, G.M., Campelo, A.K., Kattan, M.W. 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.
- Rodrigues, J., de Castro, M., Cancho, V.G. and Balakrishnan, N. (2009). COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, **139**, 3605–3611.
- Rodrigues, J., Cordeiro, G.M., Cancho, V.G. and Balakrishnan, N. (2015). Relaxed Poisson cure rate models. *Biometrical Journal*, **58**, 397–415.
- Talacko, J. (1956). Perks' distributions and their role in the theory of Wiener's stochastic variables. *Trabajos de estadística*, **7**, 159–174.
- Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J. and Stasinopoulos, D. (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, **39**, 1279–1293.

5 PREDICTING THE CURE RATE OF BREAST CANCER USING A NEW REGRESSION MODEL WITH FOUR REGRESSION STRUCTURES

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 **GAMLSS** 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, **GAMLSS**.

5.1 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 increasing, requiring a flexible statistical distributions 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 Boag (1949) Berkson and Gage (1952) and Farewell (1982). 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 (1996) 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 (Yakovlev and Tsodikov, 1996), 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 i th individual. Given M , the random variables Z_i , for $i = 1, \dots, M$, 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

$$S_p(t) = \exp[-\tau F(t)], \quad (5.1)$$

where $S_p(t)$ is the unconditional survival function of t for the entire population. Note that when $t \rightarrow \infty$, $S_p(t) \rightarrow e^{-\tau} = p$, where $0 \leq p \leq 1$ denotes the cured proportion. The probability density function (pdf) corresponding to the survival function (5.1) is given by

$$f_p(t) = \tau f(t) \exp[-\tau F(t)]. \quad (5.2)$$

Note that equation (5.2) 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 DeCastro *et al.* (2010). 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 (Hellwig *et al.*, 2010). 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 (Ramires *et al.*, 2016) was introduced to accommodate various shapes of skewness, kurtosis and bi-modality. The LSC pdf can be expressed as

$$f(t; \mu, \sigma, \nu) = \frac{\nu}{t \sigma \pi} \frac{\cosh\left(\frac{\log(t) - \mu}{\sigma}\right)}{\nu^2 \sinh^2\left(\frac{\log(t) - \mu}{\sigma}\right) + 1}, \quad (5.3)$$

where $\mu \in R$ 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 (5.3) is given by

$$F(t; \mu, \sigma, \nu) = \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh \left(\frac{\log(t) - \mu}{\sigma} \right) \right]. \quad (5.4)$$

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.* (2009) investigated local influence in generalized log-gamma regression models with cure fraction, Silva *et al.* (2008) adapted global and local influence methods in log-Burr XII regression models with censored data and Hashimoto *et al.* (2012) 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, is 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 (Rodrigues *et al.*, 2009) was introduced in the *generalized additive model for location, scale and shape* (GAMLSS) (Stasinopoulos and Rigby, 2007) package of the R software (R Core Team, 2015), 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 (DeCastro *et al.*, 2010); the standard mixture Weibull model with a frailty term was also introduced in the GAMLSS package by Calsavara *et al.* (2013), 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 5.2, 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 5.3, 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 5.4. Section 5.5 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 5.6 to illustrate the flexibility of the new regression model. Finally, we offer some conclusions in Section 5.7.

5.2 The LSCp model

Based on the LSC distribution, we define the LSCp model by inserting (5.3) and (5.4) in equation (5.2). The pdf and survival function of the LSCp model are given by

$$f_p(t; \mu, \sigma, \nu, \tau) = \frac{\tau\nu}{t\sigma\pi} \frac{\cosh(w)}{\nu^2 \sinh^2(w) + 1} \exp \left\{ -\frac{\tau}{2} - \frac{\tau}{\pi} \arctan[\nu \sinh(w)] \right\} \quad (5.5)$$

and

$$S_p(t; \mu, \sigma, \nu, \tau) = \exp \left\{ -\frac{\tau}{2} - \frac{\tau}{\pi} \arctan[\nu \sinh(w)] \right\}, \quad (5.6)$$

respectively, where $w = \frac{\log(t) - \mu}{\sigma}$, $\mu \in R$ 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 (5.5) 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

$$S(t; \mu, \sigma, \nu, \tau) = \frac{\exp \left\{ -\frac{\tau}{2} - \frac{\tau}{\pi} \arctan[\nu \sinh(w)] \right\} - \exp(-\tau)}{1 - \exp(-\tau)} \quad (5.7)$$

and

$$h_p(t; \mu, \sigma, \nu, \tau) = \frac{\tau \nu}{t \sigma \pi} \frac{\cosh(w)}{\nu^2 \sinh^2(w) + 1}. \quad (5.8)$$

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 (Li *et al.*, 2001; Ibrahim *et al.*, 2001; Cooner *et al.*, 2007). The cure model in (5.1) is identifiable if $F(\cdot)$ is a parametric model (Li *et al.*, 2001).

The functions (5.5), (5.6) and (5.8) are implemented in the R software and can be easily accessed by following the steps in the box displayed on the right. Plots of the LSCp survival and hazard functions for selected parameter values are displayed in Figures 5.1 and 5.2, respectively.

```
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
```

Figure 5.1 reveals clearly the bi-modality and symmetric effects caused by the parameters σ and ν , respectively. Further, Figure 5.2 indicates that the hrf of T has decreasing, unimodal, and bimodal shapes.

Figure 5.1. 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 5.2. 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 .

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.

5.3 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 (Ortega *et al.*, 2015). Similarly, the generalized log-gamma regression model with cure fraction (Ortega *et al.*, 2009) 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 (Rigby and Stasinopoulos, 2005) 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.

5.3.1 Definition

Let $T \sim LSCp(t; \boldsymbol{\theta})$, where $\boldsymbol{\theta}^T = (\mu, \sigma, \nu, \tau)$ denotes the vector of parameters of the pdf (5.5). Consider independent observations t_i 's conditional on the parameter vector $\boldsymbol{\theta}_i$ (for $i = 1, 2, \dots, n$) 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

$$\boldsymbol{\mu} = g_1(\mathbf{X}_1\boldsymbol{\beta}_1), \quad \boldsymbol{\sigma} = g_2(\mathbf{X}_2\boldsymbol{\beta}_2), \quad \boldsymbol{\nu} = g_3(\mathbf{X}_3\boldsymbol{\beta}_3), \quad \boldsymbol{\tau} = g_4(\mathbf{X}_4\boldsymbol{\beta}_4), \quad (5.9)$$

where $g_k(\cdot)$, for $k = 1, 2, 3, 4$, denote the injective and twice continuously differentiable monotonic link functions, $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \dots, \beta_{m_k k})^T$ is a parameter vector of length $(m_k + 1)$, m_k denotes the number of explanatory variables related to the k th 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 5.4. For the following sections, we shall consider the identity link function for $g_1(\cdot)$ and the logarithmic link function for $g_k(\cdot)$ ($k = 2, 3, 4$).

5.3.2 Inference

Consider a sample of n -independent observations $t_1, \dots, t_n$. 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 $(y_1, \delta_1, \mathbf{x}_{k1}), \dots, (y_n, \delta_n, \mathbf{x}_{kn})$, 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

$$\begin{aligned} l(\boldsymbol{\theta}) = & \sum_{i \in F} \left\{ \log(\tau_i) + \log(\nu_i) - \log(\sigma_i \pi) - \log(y_i) + \log \cosh(w_i) - \log [1 + \nu_i^2 \sinh^2(w_i)] \right\} \\ & - \sum_{i \in F} \sum_{i \in C} \tau_i \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan [\nu_i \sinh(w_i)] \right\}, \end{aligned} \quad (5.10)$$

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 (5.9) by specifying appropriate link functions for $g_k(\cdot)$, i.e., $\mu_i = \beta_{01} + \beta_{11}x_{i1} + \dots + \beta_{m_k 1}x_{im_k}$.

The numerical maximization of the log-likelihood function (5.10) can be easily performed in the GAMLSS package in R. The advantage of this package is that we can use different maximization meth-

ods. 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 (Rigby and Stasinopoulos, 2005; Stasinopoulos and Rigby, 2007). 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.

Here, we present an example of how to maximize the likelihood (5.10) in the R software. For the steps that will be presented below, consider the box on the right side. 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 (5.9). 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 (5.9), e.g., $\tau = \exp(\beta_{04})$. The fit of the LSCp model gives the vector of estimated cured proportion

```
m1=gamlss(Surv(T,D)~x1+x2,
  sigma.formula=~x1+x2,
  nu.formula=~x1+x2,
  tau.formula=~x1+x2,
  family=cens("LSCp"))
```

$$\hat{\mathbf{p}} = \exp[-\exp(\mathbf{X}_4\hat{\boldsymbol{\beta}}_4)], \quad 0 < \hat{\mathbf{p}} < 1, \quad (5.11)$$

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

The asymptotic distribution of $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ 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 $\ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}})$, i.e., the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$ given by $\ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\hat{\boldsymbol{\theta}}}$. The multivariate normal $N_m(\mathbf{0}, \ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}})^{-1})$ 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 $\tilde{\boldsymbol{\theta}}$ 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 $\Lambda = 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})]$. 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 (5.5) and (5.6), we generalize the scale parameter σ by σ_i , where the parameter σ_i can be modelled 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 $\Lambda = 2[\ell(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\beta}}_4) - \ell(\tilde{\boldsymbol{\beta}}_1, \sigma_0, \tilde{\boldsymbol{\beta}}_3, \tilde{\boldsymbol{\beta}}_4)]$, where $\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_3$ and $\tilde{\boldsymbol{\beta}}_4$ are the restricted MLEs of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$, respectively, obtained from the maximization of (5.10) under $\mathcal{H}_0 : \sigma_i = \sigma_0$. Analogously, we can perform the same tests of hypotheses for the parameters μ, ν and τ .

5.4 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.

5.4.1 Select the distribution

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 $GD = -2l(\hat{\theta})$, $AIC = GD + 2k$ and $BIC = GD + \log(n)k$, respectively, where $l(\hat{\theta})$ 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 in the box on the right.

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

5.4.2 Selecting explanatory variables

For the LSCp GAMLSS regression, the selection of the terms for all parameters is performed using a stepwise AIC procedure (Voudouris *et al.*, 2012). There are many different strategies that could be applied for selection of the terms used to model the four parameters μ , σ , ν and τ . 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 μ , with σ , ν and τ fitted as constants. After that, repeat the same procedure to select the model for σ , ν and τ , 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 ν , with μ , σ and τ fitted as constants and repeat this procedure for σ and μ , respectively. At the end of the steps described above, the final model may contain different subsets from χ for μ , σ , ν and τ .

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 box 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 .

```
m1=gamlss(Surv(T,D)~1,family=cens("LSCp"))
m2=stepGAICAll.A(m1,scope=list(lower=~1,
                                upper=~x1+x2+x3))
```

5.4.3 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 (Dunn and Smyth, 1996), which are given by $\hat{r}_i = \Phi^{-1}(\hat{u}_i)$, where $\Phi^{-1}(\cdot)$ is the qf of the standard normal variate and $\hat{u}_i = F(t_i|\hat{\theta}_i)$. For censored response variables, $\hat{u}$ is defined as a random value from a uniform distribution on the interval $[1 - S(t_i|\hat{\theta}_i), 1]$.

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) (Buuren and Fredriks, 2001).

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

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


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.

5.4.4 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 (Cook, 1986), 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 (Cook and Weisberg, 1982), 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 (5.9), the log-likelihood function (5.10) for θ is denoted by $l(\theta)$. Let $\hat{\theta}_{(-i)}^T = (\hat{\mu}_{(-i)}^T, \hat{\sigma}_{(-i)}^T, \hat{\nu}_{(-i)}^T, \hat{\tau}_{(-i)}^T)$ be the MLEs of μ , σ , ν and τ obtained from $l(\theta_{(-i)})$. To assess the influence of the i th case on the MLE $\hat{\theta}$ the idea is to compare the difference between $\hat{\theta}_{(-i)}$ and $\hat{\theta}$. If deletion of a case seriously influences the estimates, more attention should be given to that case. Hence, if $\hat{\theta}_{(-i)}$ is far from $\hat{\theta}$, then the i th case is regarded as an influential observation. A popular measure of the difference between $\hat{\theta}_{(-i)}$ and $\hat{\theta}$, called log-likelihood distance, is given by

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

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 $\hat{\theta}$ as initial vector to obtain the MLE $\hat{\theta}_{(-i)}$. An example of how to calculate $LD_i(\theta)$ using the **GAMLSS** package is given in supplementary material.

5.5 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 (5.11) and the censored failure times percentage is denoted by ψ .

We can simulate LSCp random variables using the quantile function (qf), which is obtained by inverting $F(t) = 1 - S(t) = u$, where $S(t)$ represents the survival function for non-censored observations (5.7). The qf of $T \sim \text{LSCp}(t, \mu, \sigma, \nu, \tau)$ is given by

$$T = Q(u) = \exp \left(\mu + \sigma \operatorname{arcsinh} \left\{ \frac{1}{\nu} \tan [\pi (k(u, \tau) - 0.5)] \right\} \right), \quad (5.12)$$

where $k(u, \tau) = -\log[(u - 1)(e^{-\tau} - 1)]$. Equation (5.12) 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.

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 $n_c = ne^{-\tau}$, and by the sample of the observed times T , with size $n_t = n - n_c$.

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

Now, we generate n_t observations using (5.12) 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 in the box 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 supplementary material.

Using equation (5.9), 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 1,000 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 in the box presented above are used for maximizing the total log-likelihood function (5.10).

The results are reported in Table 5.1 and, for a visual analysis, we present in Figure 5.3 the generated and the estimated (considering the AEs given in Table 5.1) survival functions for $n = 50, 100$ and 150 and considering the two groups represented by the explanatory variable x_{1i} .

Table 5.1. The AEs, biases and MSEs based on 1,000 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
	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
	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
	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
	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

Figure 5.3. LSCp survival functions at the true parameter values and at the AEs obtained in Table 5.1 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 of the Monte Carlo study in Table 5.1 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 5.3.

5.6 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 (DeSantis *et al.*, 2015). One in 8 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 (Fitzgibbons *et al.*, 2000). Due to the fact of the introduction of new imaging modalities, the multifocality has also been considered as an 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 (Berg *et al.*, 2004). 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 (Kattan *et al.*, 2004). 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 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 ($i = 1, \dots, 284$) 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 (5.9) disregarding regression variables. Table 5.2 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 (5.11), 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 (Yakovlev and Tsodikov, 1996). Then, the patients can be considered as cured. Figure 5.4 provides the plots of the estimated and empirical survival function. Table 5.2 and Figure 5.4 indicate that the LSCp model provides a good fit to these data.

Table 5.2. 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

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 regression models were fitted to these data (Ortega *et al.*, 2015) 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.,

$$\log \tau = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \beta_4 \mathbf{X}_4 + \beta_5 \mathbf{X}_5 + \beta_6 \mathbf{X}_6.$$

The values of the GD, AIC and BIC statistics for the fitted models are listed in Table 5.3. 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.

Figure 5.4. The estimated and empirical survival functions.

Table 5.3. 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

Using the steps described in Section 5.4 to select the additive terms for the different parameters, we present results for the model parameters defined by

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

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 supplementary material for future research. The additive terms selected for the Weibullcr model are

$$\begin{aligned}\mu_i &= \exp(\beta_{01} + \beta_{41}x_{i4} + \beta_{51}x_{i5}), \quad \sigma_i = \exp(\beta_{02}) \quad \text{and} \\ \nu_i &= \text{logit}(\beta_{03} + \beta_{23}x_{i2} + \beta_{33}x_{i3} + \beta_{43}x_{i4} + \beta_{53}x_{i5} + \beta_{63}x_{i6}).\end{aligned}$$

Table 5.4 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 5.4. The MLEs, corresponding SEs and p -values of the estimates from the fitted LSCp and Weibullcr GAMLSS regression.

Model	Parameter	Estimate	SE	p -value	Parameter	Estimate	SE	p -value
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

Table 5.5 provides the formal tests to verify the significance of the explanatory variables presented in Table 5.4 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.

Table 5.5. LR tests

Parameter	$l(\theta)$	Λ	p -value	Parameter	$l(\theta)$	Λ	p -value
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

The criteria obtained for the fitted models in Table 5.4 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 5.5 in order to verify the adequacy and the assumptions of the fitted models. In Figures 5.5(a)-(b) we note that the quantile residuals have an approximately normal distribution. The WP given in Figure 5.5(c) reveals that the proposed regressions for modeling the mean, variance, skewness and kurtosis are correct. Figures 5.5(d)-(e) indicate that the Weibullcr model does not present a good fit for extreme values. Also, in Figure 5.5(f) we can note a U-shape in the WP, thus indicating failure for modelling 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 (5.11), the estimated cured proportions can be determined using the results obtained in (5.4) as $p_i = \exp[-\exp(-4.290 + 2.817x_{i4} + 1.195x_{i6} + 0.288x_{i3})]$. In Figure 5.6, 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 5.6) and the low-risk g_2 group composed by $X_4 = 0$, and $X_6 = 0$ (black line in Figure 5.6). In Figure 5.7, we

Figure 5.5. Residual analysis: For the LSCp and Weibuller models, (a)-(d) Density of the quantile residuals, (b)-(e) Q-Q plot and (c)-(f) WP, respectively.

Figure 5.6. The estimated cured proportions for each level of X_4 and X_6 for all range of X_3 .

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 $LD_i(\theta)$. Figure 5.8 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 5.7(b) pink curve).

5.7 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 modelled as parametric functions of explanatory variables. Procedures for

Figure 5.7. 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 .

Figure 5.8. Index plots for $|LD_i(\theta)|$.

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.

5.8 Supplementary material

5.8.1 Codes used in global influence

```
##### Final model for breast cancer #####
```



```

# Let t the survival times and censur the failure indicator
m2=gamlss(Surv(t,censur)~x4,sigma.fo=~x2+x6,nu.fo=~x5,tau.fo=~x3+x6+x4,family=cens("LSCp"))
v1=as.numeric(logLik(m2))#likelihood

##### LD Analysis #####
vtot=c()
for(i in 1:length(t)){
mod=gamlss(Surv(t[-i],censur[-i]) ~x4[-i],sigma.fo=~x2[-i]+x6[-i],nu.fo=~x5[-i],
tau.fo=~x3[-i]+x6[-i]+x4[-i],family=cens("LSCp"),c.crit=.1,
mu.start = m2$mu.fv[-i], sigma.start = m2$sigma.fv[-i],
nu.start = m2$nu.fv[-i],tau.start = m2$tau.fv[-i])
v=as.numeric(logLik(mod)); vtot=c(vtot,v) }
vcomp=c(rep(v1,length(vtot)))
LDp1=(2*(vcomp-vtot))
coll=ifelse(censur==0,"gray0","dimgray")
plot(abs(LDp1),pch=16,ylab="|Likelihood distance|",type="h",lwd=2,col=coll)

```

5.8.2 Codes used in simulation study

```

#####First simulation#####
mu1=mu2=sigma1=sigma2=nu1=nu2=tau1=tau2=c()
for(n in c(25,50,75)){for(i in 1:1000){

#random values
rLSCp(n,1.5,0.3,0.1,2) ; t1=Times;c1=Delta
rLSCp(n,2.5,0.2,0.5,1); t2=Times;c2=Delta
time=c(t1,t2);censur=c(c1,c2); grup=c(rep("a",n),rep("b",n))

m1=gamlss(Surv(time,censur)~grup, sigma.fo=~grup,nu.fo=~grup,tau.fo=~grup,
family=cens(LSCp),n.cyc=300,c.crit=.01)

mu1=c(mu1,m1$mu.fv[1]);mu2=c(mu2,m1$mu.fv[n+1])
sigma1=c(sigma1,m1$sigma.fv[1]);sigma2=c(sigma2,m1$sigma.fv[n+1])
nu1=c(nu1,m1$nu.fv[1]);nu2=c(nu2,m1$nu.fv[n+1])
tau1=c(tau1,m1$tau.fv[1]);tau2=c(tau2,m1$tau.fv[n+1])}}

a=1:1000;b=1001:2000;c=2001:3000
AE=c(mean(mu1[a]),mean(mu2[a]),mean(sigma1[a]),mean(sigma2[a]),mean(nu1[a]),
mean(nu2[a]),mean(tau1[a]),mean(tau2[a]),mean(mu1[b]),mean(mu2[b]),
mean(sigma1[b]),mean(sigma2[b]),mean(nu1[b]),mean(nu2[b]),
mean(tau1[b]),mean(tau2[b]),mean(mu1[c]),mean(mu2[c]),mean(sigma1[c]),
mean(sigma2[c]),mean(nu1[c]),mean(nu2[c]),mean(tau1[c]),mean(tau2[c]))

Bias=abs(AE-rep(c(1.5,2.5,0.3,0.2,0.1,0.5,2,1),3))
MSE=c(var(mu1[a]),var(mu2[a]),var(sigma1[a]),var(sigma2[a]),var(nu1[a]),var(nu2[a]),
var(tau1[a]),var(tau2[a]),var(mu1[b]),var(mu2[b]),var(sigma1[b]),var(sigma2[b]),
var(nu1[b]),var(nu2[b]),var(tau1[b]),var(tau2[b]),var(mu1[c]),var(mu2[c]),
var(sigma1[c]),var(sigma2[c]),var(nu1[c]),var(nu2[c]),var(tau1[c]),var(tau2[c]))

#####Second simulation#####
mu1=mu2=sigma1=sigma2=nu1=nu2=tau1=tau2=c()
for(n in c(25,50,75)){ for(i in 1:1000){

#random values
rLSCp(n,1.5,0.3,0.1,2) ; t1=Times;c1=Delta
rLSCp(n,2.5,0.2,0.5,1); t2=Times;c2=Delta
t1=t1[order(t1)];t2=t2[order(t2)]; c1=c1[order(t1)];c2=c2[order(t2)];
r1=round(sum(c1)*0.1,0) #number of lifetime censored
r2=round(sum(c2)*0.1,0) #number of lifetime censored
s1=sum(c1);s2=sum(c2) ; rand1=runif(s1);rand2=runif(s2)
c1[order(rand1)[1:r1]]=0; c2[order(rand2)[1:r2]]=0
time=c(t1,t2); censur=c(c1,c2); grup=c(rep("a",n),rep("b",n))

m1=gamlss(Surv(time,censur)~grup, sigma.fo=~grup,nu.fo=~grup,tau.fo=~grup,
family=cens(LSCp),n.cyc=300,c.crit=.01)

mu1=c(mu1,m1$mu.fv[1]);mu2=c(mu2,m1$mu.fv[n+1])
sigma1=c(sigma1,m1$sigma.fv[1]);sigma2=c(sigma2,m1$sigma.fv[n+1])
nu1=c(nu1,m1$nu.fv[1]);nu2=c(nu2,m1$nu.fv[n+1])
tau1=c(tau1,m1$tau.fv[1]);tau2=c(tau2,m1$tau.fv[n+1])}}

a=1:1000;b=1001:2000;c=2001:3000

```

```

AE=c(mean(mu1[a]),mean(mu2[a]),mean(sigma1[a]),mean(sigma2[a]),mean(nu1[a]),mean(nu2[a]),
      mean(tau1[a]),mean(tau2[a]),mean(mu1[b]),mean(mu2[b]),mean(sigma1[b]),mean(sigma2[b]),
      mean(nu1[b]),mean(nu2[b]),mean(tau1[b]),mean(tau2[b]),mean(mu1[c]),mean(mu2[c]),
      mean(sigma1[c]),mean(sigma2[c]),mean(nu1[c]),mean(nu2[c]),mean(tau1[c]),mean(tau2[c]))

Bias=abs(means-rep(c(1.5,2.5,0.3,0.2,0.1,0.5,2,1),3))
MSE=c(var(mu1[a]),var(mu2[a]),var(sigma1[a]),var(sigma2[a]),var(nu1[a]),var(nu2[a]),var(tau1[a]),
      var(tau2[a]),var(mu1[b]),var(mu2[b]),var(sigma1[b]),var(sigma2[b]),var(nu1[b]),var(nu2[b]),
      var(tau1[b]),var(tau2[b]),var(mu1[c]),var(mu2[c]),var(sigma1[c]),var(sigma2[c]),var(nu1[c]),
      var(nu2[c]),var(tau1[c]),var(tau2[c]))

```

5.8.3 Codes of the Weibullcr GAMLSS

```

source("https://goo.gl/TSNomS") #codes implemented in the GAMLSS
dweibull(x,mu,sigma,nu) #pdf
pweibull(x,mu,sigma,nu) #cdf

```

References

- Berg, WA, Gutierrez L, NessAiver MS, Carter WB, Bhargavan M, Lewis RS and Ioffe OB. (2004). Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer 1. *Radiology*, 233: 830–849.
- Berkson J and Gage RP. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47: 501–515.
- Boag JW. (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 SV and Fredriks M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, 20: 1259–1277.
- Calsavara VF, Tomazella VL and Fogo JC. (2013). The effect of frailty term in the standard mixture model. *Chilean Journal of Statistics*, 4: 95–109.
- de Castro M, Cancho VG and Rodrigues J. (2010). A hands-on approach for fitting long-term survival models under the GAMLSS framework. *Computer methods and programs in biomedicine*, 97: 168–177.
- DeSantis CE, Bray F, Ferlay J, Lortet-Tieulent J, Anderson, BO and Jemal A. (2015). International variation in female breast cancer incidence and mortality rates. *Cancer Epidemiology Biomarkers & Prevention*, 24: 1495–1506.
- Cook RD. (1986). Assessment of local influence. *Journal of the Royal Statistical Society B*, 48: 133–169.
- Cook RD. and Weisberg S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.
- Cooner F, Banerjee S, Carlin BP and Sinha D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association*, 102: 560–572.
- Dunn PK and Smyth GK. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5: 236–244.
- Farewell VT. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38: 1041–1046.
- Fitzgibbons PL, Page DL, Weaver D, Thor AD, Allred DC, Clark GM, et al. (2000). Prognostic factors in breast cancer: College of American Pathologists consensus statement 1999. *Archives of pathology & laboratory medicine*, 124: 966–978.

- Hashimoto EM, Ortega EMM, Cordeiro GM and Barreto ML. (2012). The Log-Burr XII regression model for grouped survival data. *Journal of biopharmaceutical statistics*, 22: 141–159.
- Hellwig B, Hengstler JG, Schmidt M, Gehrmann MC, Schormann W and Rahnenfuhrer J. (2010). Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC bioinformatics*, 11: 1.
- Ibrahim JG, Chen MH and Sinha D. (2001). *Bayesian Survival Analysis*. Springer: New York.
- Kattan WM, Giri D, Panageas KS, Hummer A, Cranor M, Zee KJV, Hudis CA, Norton L, Borgen PI and Tan LK. (2004). A tool for predicting breast carcinoma mortality in women who do not receive adjuvant therapy. *Cancer*, 101: 2509–2515 .
- Li CS, Taylor JM and Sy JP. Identifiability of cure models. (2001). *Statistics & Probability Letters*, 54: 389–395.
- Ortega EM, Cancho VG and Paula GA. (2009). Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis*, 15: 79–106.
- Ortega EM, Cordeiro GM, Campelo AK, Kattan MW and Cancho VG. (2015). A power series beta Weibull regression model for predicting breast carcinoma. *Statistics in Medicine*, 34: 1366–1388.
- R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- 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 RA and Stasinopoulos DM. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54: 507–554.
- Rodrigues J, de Castro M, Cancho VG and Balakrishnan N. (2009). COM–Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning Inference*, 139: 3605–3611.
- Silva GO, Ortega EMM, Cancho VG and Barreto ML. (2008). Log-Burr XII regression models with censored data. *Computational Statistics & Data Analysis*, 52: 3820–3842.
- Stasinopoulos DM and Rigby, RA. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23: 1–46.
- 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.
- Yakovlev A and Tsodikov AD. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications. Mathematical Biology and Medicine, Vol. 1*. World Scientific, New Jersey.

6 A FLEXIBLE SEMIPARAMETRIC REGRESSION MODEL FOR BIMODAL, ASYMMETRIC AND CENSORED DATA

Abstract: In this paper, we propose a new semiparametric heteroscedastic regression model allowing for positive and negative skewness and bimodal shapes using the B-spline basis for nonlinear effects. The proposed distribution is based on the generalized additive models for location, scale and shape framework in order to model any or all the parameters of the distribution using parametric linear and/or nonparametric smooth functions of explanatory variables. We motivate the new model by means of Monte Carlo simulations, thus ignoring the skewness and bimodality of the random errors in semiparametric regression models. We may introduce biases on the parameter estimates and/or on the estimation of the associated variability measures. An iterative estimation process and some diagnostic methods are investigated. Applications to two real data sets are presented and the method is compared to the usual regression methods.

Keywords: GAMLSS; global influence; P-splines; residual analysis.

6.1 Introduction

Nonlinear regression models are commonly applied in areas such as biology, chemistry, medicine, economics and engineering. The analysis based on models under normal errors and constant variance is most popular when the variable of interest is continuous due to desirable statistical properties and a comprehensive theory. However, if the random error distribution happens to be non-normal, in particular, if it has heavier-than-normal tails or bimodal characteristics, then the accuracy of the ordinary least squares solutions is lost, introducing biases on the parameter estimates. For more accurate models, a large number of new parametric and semiparametric models to extend well-known distributions and to provide flexibility in modeling data have been investigated in the last years. Recently, Vanegas and Paula (2015) proposed a semiparametric regression model in which the distribution of the response is asymmetric (see also Vanegas and Paula, 2016); Cancho *et al.* (2010) studied nonlinear skew-normal regression models using classical and Bayesian approaches; Xu *et al.* (2015) proposed the skew-normal semiparametric model, which provides a useful extension of the normal regression model. In other words, a standard assumption in linear or nonlinear regression analysis is homogeneity of the error variances. Violation of this assumption can have adverse consequences for the efficiency of the estimators. So, it is important to check for heteroscedasticity whenever it is considered a possibility (Cysneiros *et al.*, 2010). In this sense, Lachos *et al.* (2011) introduced heteroscedastic nonlinear regression models based on scale mixtures of skew-normal distributions; Voudouris *et al.* (2012) showed an application of the Box-Cox power exponential distribution modeling the location, scale and skewness parameters using P-splines bases; and Nakamura *et al.* (2016) introduced the Birnbaum-Saunders power distribution modeling its parameters using smooth functions.

Although the models studied in these papers are attractive, they have several limitations. Most of the proposed models are not able to capture the presence of bimodality and negative skewness of the random errors. As an alternative, for modeling a lifetime $T > 0$, Ramires *et al.* (2016) introduced the exponentiated log-sinh Cauchy (ELSC) distribution to accommodate various shapes of skewness, kurtosis and bi-modality. Based on the logarithm $Y = \log(T)$, where T has the ELSC distribution, we defined the *exponentiated sinh Cauchy* (ESC) linear regression model in the generalized additive model for location, scale and shape framework (Rigby and Stasinopoulos, 2005), where all parameters are modeled by explanatory variables. The ESC regression model proved to be very flexible to fit data with modal

and bimodal shapes as well as positive and negative skewness. The probability density function (pdf) and cumulative distribution function (cdf) of the ESC distribution are given by

$$f(y; \mu, \sigma, \nu, \tau) = \frac{\tau\nu}{\sigma\pi} \frac{\cosh\left(\frac{y-\mu}{\sigma}\right)}{\nu^2 \sinh^2\left(\frac{y-\mu}{\sigma}\right) + 1} \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh \left(\frac{y-\mu}{\sigma} \right) \right] \right\}^{\tau-1} \quad (6.1)$$

and

$$F(y; \mu, \sigma, \nu, \tau) = \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh \left(\frac{y-\mu}{\sigma} \right) \right] \right\}^{\tau}, \quad (6.2)$$

respectively, where $\mu \in R$ 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 skewness parameter. The ESC density (6.1) was originally introduced and studied by Cooray (2013), disregarding the regression structure, to modeling symmetric, right and left skewed and bimodal data sets.

We propose a general class of semiparametric ESC regression models using P-splines in the additive terms. The sections are organized as follows. In Section 6.2, we define the ESC GAMLSS semiparametric regression model. We also discuss inferential issues, smooth function, methods for generating random values, residual analysis, model selection strategies and global influence measure. In Section 6.3, we perform some Monte Carlo simulations on the finite sample behavior of the maximum likelihood estimates (MLEs). Applications to two real data sets are presented in Section 6.4, which illustrate the flexibility of the proposed class of regression models. Finally, we offer some conclusions in Section 6.6.

6.2 The ESC regression model

In many practical applications, the response variables are affected by explanatory variables. In the presence of explanatory variables with nonlinear effects, semiparametric models are widely used and when their models provide a good fit, they tend to give more precise estimates of the quantities of interest. Recently, several regression models have been proposed in the literature by considering the class of location models. For example, Ramires *et al.* (2013) proposed the log-beta generalized half-normal geometric regression model for censored data, Cordeiro *et al.* (2015) presented the log-generalized Weibull-log-logistic regression model for predicting longevity of the mediterranean fruit fly and Ortega *et al.* (2015) studied a power series beta Weibull regression model for predicting breast carcinoma. A disadvantage of the class of the location models is that the variance, skewness, bimodality, kurtosis and other parameters are not modelled explicitly in terms of the explanatory variables but implicitly through their dependence on the location parameter. As an alternative, the *generalized additive model for location, scale and shape* (GAMLSS) (Rigby and Stasinopoulos, 2005), wherein the systematic part of the model is expanded, allows not only the location but all parameters of the conditional distribution of Y to be modelled as parametric functions of explanatory variables.

6.2.1 Definition

Let $\theta^T = (\mu, \sigma, \nu, \tau)$ denote the vector of parameters of the pdf (6.1). We consider independent observations y_i conditional on θ_i (for $i = 1, 2, \dots, n$), with pdf $f(y_i; \theta_i)$, where $\theta_i^T = (\mu_i, \sigma_i, \nu_i, \tau_i)$ is a vector of parameters related to the response variable. The ESC linear regression model, linking the response variable y_i and the explanatory variables, can be defined by

$$y_i = \mu_i + \sigma_i z_i, \quad i = 1, \dots, n, \quad (6.3)$$

where the random error $Z_i = (Y_i - \mu_i)/\sigma_i$ has pdf given by

$$f(z; \nu, \tau) = \frac{\tau\nu}{\pi} \frac{\cosh(z)}{\nu^2 \sinh^2(z) + 1} \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan [\nu \sinh(z)] \right\}^{\tau-1}, \quad \text{for } z \in R. \quad (6.4)$$

Plots of the density function (6.4) for selected parameter values are displayed in Figure 6.1. We can note that the proposed model is able to fit data with modal and bimodal shapes as well as positive and negative skewness.

Figure 6.1. Plots of the density function (6.4) for several values of τ : (a) $\nu = 0.3$; (b) $\nu = 0.8$.

We can define the vector of parameters θ using appropriate link functions as

$$\theta = \begin{bmatrix} \mu \\ \sigma \\ \nu \\ \tau \end{bmatrix} = \begin{bmatrix} g_1(\mathbf{X}_1\beta_1) \\ g_2(\mathbf{X}_2\beta_2) \\ g_3(\mathbf{X}_3\beta_3) \\ g_4(\mathbf{X}_4\beta_4) \end{bmatrix} \text{ or } \theta_i = \begin{bmatrix} \mu_i \\ \sigma_i \\ \nu_i \\ \tau_i \end{bmatrix} = \begin{bmatrix} g_1(\beta_{01} + \mathbf{X}_1[i, 2]\beta_{11} + \dots + \mathbf{X}_1[i, p_1 + 1]\beta_{p_11}) \\ g_2(\beta_{02} + \mathbf{X}_2[i, 2]\beta_{12} + \dots + \mathbf{X}_2[i, p_2 + 1]\beta_{p_22}) \\ g_3(\beta_{03} + \mathbf{X}_3[i, 2]\beta_{13} + \dots + \mathbf{X}_3[i, p_3 + 1]\beta_{p_33}) \\ g_4(\beta_{04} + \mathbf{X}_4[i, 2]\beta_{14} + \dots + \mathbf{X}_4[i, p_4 + 1]\beta_{p_44}) \end{bmatrix}, \quad (6.5)$$

where p_k represents the number of explanatory variables related to the k th parameter, $g_1(\cdot)$ is an injective and twice continuously differentiable functions, $g_k(\cdot)$, for $k = 2, 3, 4$, are known positive continuously differentiable function containing values of the explanatory variables, $\beta_k = (\beta_{0k}, \beta_{1k}, \dots, \beta_{p_kk})^T$ is a parameter vector of length $(p_k + 1)$ and $\mathbf{X}_k$ is a known model matrix of order $n \times (p_k + 1)$, whose elements are given by $\mathbf{X}_k[i, p_k]$. The total number of parameters to be estimated is defined by $p = p_1 + p_2 + p_3 + p_4 + 4$. In the following sections, we will consider the identity link function for $g_1(\cdot)$ and the logarithmic link function for $g_k(\cdot)$ for $k = 2, 3, 4$. The GAMLSS framework family extends two major classes of regression models. The class of location models follows by taking $p_2 = p_3 = p_4 = 0$. For $p_3 = p_4 = 0$, $p_1 \neq 0$ and $p_2 \neq 0$, we obtain the regression model with heteroscedastic errors, which can be used as an alternative to transform the response variable. However, the choice of parameters to be modeled by explanatory variables will depend on the data set.

6.2.2 Nonparametric additive functions

The ESC GAMLSS model allows the user to model the distribution parameters μ , σ , ν and τ as linear, nonlinear parametric, nonparametric (smooth) function of the explanatory variables and/or random-effects terms. The parametric regression structure (6.5) can be extended to semiparametric structure as

$$\theta = \begin{bmatrix} \mu \\ \sigma \\ \nu \\ \tau \end{bmatrix} = \begin{bmatrix} g_1\left(\mathbf{X}_1\beta_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1})\right) \\ g_2\left(\mathbf{X}_2\beta_2 + \sum_{j=2}^{J_2} h_{j2}(\mathbf{x}_{j2})\right) \\ g_3\left(\mathbf{X}_3\beta_3 + \sum_{j=3}^{J_3} h_{j3}(\mathbf{x}_{j3})\right) \\ g_4\left(\mathbf{X}_4\beta_4 + \sum_{j=4}^{J_4} h_{j4}(\mathbf{x}_{j4})\right) \end{bmatrix}, \quad (6.6)$$

where $h_{jk}(\mathbf{x}_{jk})$ are smooth functions of the explanatory variables $\mathbf{x}_{jk}$ for $k = 1, 2, 3, 4$ and $j = 1, \dots, J_k$. The explanatory variables can be similar or different for each of the distribution parameters, which can be considered as linear functions, may be represented by smooth functions or both.

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 Stasinopoulou (2005) proved that each smoothing function $h_{jk}(\cdot)$ can be expressed as a random effects model, i.e., $h_{jk}(\mathbf{Z}_{jk}) = \mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}$, 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). Details of the number of knots as well as the degrees of freedom can be found in Eilers and Marx (1996).

6.2.3 Estimation

In this section, we present and discuss estimation methods for three types of models. First, for the ESC parametric regression model, only parametric additive terms are taken as functions of the explanatory variables. In the second, we consider the ESC parametric regression model for censored observations. For the third model, parametric and nonparametric functions are considered for the explanatory variables. The numerical maximization of the log-likelihoods presented below can be performed in the GAMLSS package of the R software using the computational codes implemented by the first author and available at <https://goo.gl/hAIcBF>. The maximization algorithms used are the RS and CG procedures, described by Rigby and Stasinopoulou (2005) and Stasinopoulou and Rigby (2007) and available in the documentation of the GAMLSS package.

- **Parametric model**

Consider a sample of n -independent observations $y_1, \dots, y_n$. For the parametric ESC regression model (6.5), the log-likelihood for the model parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})^T$ reduces to

$$l(\boldsymbol{\theta}) = \sum_{i=1}^n \left(-\log \left[1 + \nu_i^2 \sinh^2 \left(\frac{y_i - \mu_i}{\sigma_i} \right) \right] + \log \cosh \left(\frac{y_i - \mu_i}{\sigma_i} \right) + \log(\tau_i \nu_i) - \log(\sigma_i \pi) \right. \\ \left. + (\tau_i - 1) \log \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu_i \sinh \left(\frac{y_i - \mu_i}{\sigma_i} \right) \right] \right\} \right). \quad (6.7)$$

- **Survival model**

The log-likelihood (6.7) can be easily extended to survival analysis models. Consider noninformative censoring and that the observed lifetimes and censoring times are independent. Let F and C be the sets of individuals for which y_i is the log-lifetime or log-censoring, respectively. The total log-likelihood has the form

$$l(\boldsymbol{\theta}) = \sum_{i \in F} \log f(y_i; \boldsymbol{\theta}_i) + \sum_{i \in C} \log S(y_i; \boldsymbol{\theta}_i), \quad (6.8)$$

where $\sum_{i \in F} \log f(y_i; \boldsymbol{\theta}_i)$ can be obtained using (6.7) by considering only the uncensored observations and $\sum_{i \in C} \log S(y_i; \boldsymbol{\theta}_i)$ is given by

$$\sum_{i \in C} \log S(y_i; \boldsymbol{\theta}_i) = \sum_{i \in C} \log \left(1 - \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan [\nu_i \sinh(z_i)] \right\}^{\tau_i} \right).$$

The log-likelihood (6.8) can also be maximized in the GAMLSS package using the additional package `gamlss.cens` to determine numerically the observed information corresponding to the censored observations.

- **Semiparametric model**

Considering the semiparametric model (6.6), for fixed smoothing parameters λ_{jk} , the fixed and random effects β and γ , respectively, are estimated by maximizing a penalized log-likelihood function

$$l_p = l(\theta) - \frac{1}{2} \sum_{k=1}^4 \sum_{j=1}^{J_k} \lambda_{jk} \gamma_{jk}^T \mathbf{P}_{jk} \gamma_{jk}, \quad (6.9)$$

where $l(\theta)$ is the log-likelihood function (6.7) or (6.8) and $\mathbf{P}_{jk}$ is a symmetric matrix which may depend on a vector of smoothing parameters (see Rigby and Stasinopoulos, 2005). The score functions relative to the likelihood (6.9) are given by

$$\mathbf{U}^T(\theta) = \frac{\partial l_p}{\partial \theta} = [\mathbf{U}_{\beta_1}, \mathbf{U}_{\gamma_{j1}}, \mathbf{U}_{\beta_2}, \mathbf{U}_{\gamma_{j2}}, \mathbf{U}_{\beta_3}, \mathbf{U}_{\gamma_{j3}}, \mathbf{U}_{\beta_4}, \mathbf{U}_{\gamma_{j4}}],$$

where the elements are given in Appendix A. For each smoothing term selected, and any of the parameters of the ESC distribution, there is one smoothing parameter λ associated with it. The smoothing parameters can be fixed or estimated from the data. We adopted the PQL method, described by Lee *et al.* (2006), to estimate the smoothing parameters as well as the degrees of freedom of the P-spline smooth functions. This method is implemented in the R software in the function `pb(.)` (Rigby and Stasinopoulos, 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.

6.2.4 Model strategy

In this section, we discuss different methods to select the appropriate distribution for the response variable as well as the explanatory variables to compose the regression models.

- **Select the distribution**

The selection of the appropriate distribution is performed in two stages, the fitting stage and the diagnostic stage (Section 6.2.6). In the first stage, the generalized Akaike information criterion (GAIC) is used to assess different fitted models. The GAIC is defined by $\text{GAIC}(k) = GD + k \times df$, where GD represents the global deviance given by $GD = -2l(\hat{\theta})$, $l(\hat{\theta})$ is the total log-likelihood function, df is the total effective degrees of freedom of the fitted model and k is a constant. The model with the smallest value of the criterion $\text{GAIC}(k)$ is then selected. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are special cases of the $\text{GAIC}(k)$ statistic corresponding to $k = 2$ and $k = \log(n)$, respectively.

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 calculated using the `edfAll()` function in the R software.

- **Selecting explanatory variables**

For the ESC GAMLSS model, the selection of the terms for all the parameters is done using the stepwise 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:

1. use a backward selection procedure to select an appropriate model for μ with σ , ν and τ fitted as constants;
2. use a forward selection procedure to select an appropriate model for σ given the model for μ obtained in (1) and for ν and τ fitted as constants;
3. use a forward selection procedure to select an appropriate model for τ given the model for μ and σ obtained in (2) with ν fitted as a constant;
4. use a forward selection procedure to select an appropriate model for ν given the model for μ , σ and τ obtained in (3);
5. use a backward selection procedure to select an appropriate model for τ given the model for μ , σ and ν obtained in (4);
6. use a backward selection procedure to select an appropriate model for σ , given the model for μ , ν and τ obtained (5);
7. use a backward selection procedure to select an appropriate model for μ given the model for σ , ν and τ obtained in (6).

At the end of the steps described above, the final model may contain different subsets from χ for μ , σ , ν and τ .

6.2.5 Simulation

Let a random variable Y have pdf (6.1). Inverting $F(y) = u$ in (6.2), we obtain the quantile function (qf) for Y given by

$$Q_Y(u) = \mu + \sigma \operatorname{arcsinh} \left\{ \frac{1}{\nu} \tan \left[\pi \left(u^{1/\tau} - 0.5 \right) \right] \right\}. \quad (6.10)$$

Equation (6.10) can be used for simulating random variables $y_i \sim \operatorname{ESC}(\mu, \sigma, \nu, \tau)$ by fixing μ , σ , ν and τ and setting u as a uniform random variable in the interval $(0, 1)$. We can simulate the regression models setting the parameters using the parametric (6.5) or semiparametric (6.6) structure.

6.2.6 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 in the normalized randomized quantile residuals (Dunn and Smyth, 1996), which are given by $\hat{r}_i = \Phi^{-1}(u_i)$, where $\Phi^{-1}(\cdot)$ is the qf of the standard normal variate and $u_i = F(y_i|\hat{\theta}_i)$.

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 fit adequately 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.

Finally, the fitted centile curves and the fitted conditional distribution for different values of the explanatory variable can be used to verify the goodness of fit of the model. The fitted centile curves, defined by $F(Y \leq y_u) = u$, can be easily evaluated using (6.10), where y_u is the exact $100 \times u$ centile of Y . To construct the fitted conditional distribution for different values of the explanatory variable, we use the smoothed scatterplot diagram available in the `gamlss.util` package of the R software.

6.2.7 Global influence

Since regression models are sensitive to the underlying model assumptions, performing a sensitivity analysis is strongly advisable. Cook (1986) used this idea to motivate the assessment of influence analysis. He suggested 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 (Cook and Weisberg, 1982), in which the effects or perturbations of completely removing cases from the analysis are studied. The case-deletion model for model (6.6) is given by

$$y_l = \mu_l + \sigma_l z_l, \quad l = 1, \dots, n, \quad l \neq i, \quad (6.11)$$

where the random error Z_l has a density function $f(z_l; \nu_l, \tau_l)$ given in (6.4).

In the following, a quantity with subscript “ $(-i)$ ” refers to the original quantity with the i th case deleted. For model (6.11), the log-likelihood function for θ is denoted by $l_{(-i)}(\theta)$. Let $\hat{\theta}_{(-i)}^T = (\hat{\mu}_{(-i)}^T, \hat{\sigma}_{(-i)}^T, \hat{\nu}_{(-i)}^T, \hat{\tau}_{(-i)}^T)$ be the MLEs of μ, σ, ν and τ from $l_{(-i)}(\theta)$. To assess the influence of the i th case on the MLE $\hat{\theta}^T = (\hat{\mu}^T, \hat{\sigma}^T, \hat{\nu}^T, \hat{\tau}^T)$, the basic idea is to compare the difference between $\hat{\theta}_{(-i)}$ and $\hat{\theta}$. If deletion of a case seriously influences the estimates, by changing the inference, more attention should be given to that case. Hence, if $\hat{\theta}_{(-i)}$ is far from $\hat{\theta}$, then the i th case is regarded as an influential observation. We work with a popular measure of the difference between $\hat{\theta}_{(-i)}$ and $\hat{\theta}$ given by the log-likelihood distance

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

where $l(\hat{\theta})$ is given by (6.7) for parametric models and (6.9) for semiparametric models. Note that for a specific data set and model, the penalized likelihood can potentially have multiple local maxima, so we suggest to use the MLE $\hat{\theta}$ as initial values to obtain the MLE $\hat{\theta}_{(-i)}$.

6.3 Simulation Study

We conduct a Monte Carlo simulation study under three scenarios to assess the finite sample behavior of the MLEs of the parameters for different sample sizes n . For all scenarios, we consider model (6.6), where the location and scale parameters are given by $\mu = 26 \sin(\pi x_1) + 6x_2 + 3x_3$ and $\sigma = 4$, and the variables X_1, X_2 and X_3 are generated from the uniform $[0, 2]$, binomial($n, 0.5$) and standard normal distributions, respectively. Plots of the densities of the random errors for each scenario are displayed in Figure 6.2, where the configurations are given by:

Scenario 1 bimodal symmetric density $Z \sim ESC(z, \nu = 0.05, \tau = 1)$;

Scenario 2 unimodal density with positive skewness $Z \sim ESC(z, \nu = 1.5, \tau = 5)$;

Scenario 3 unimodal density with negative skewness $Z \sim ESC(z, \nu = 1, \tau = 0.5)$.

For each scenario, the sample sizes are generated by taking $n = 50, 150$ and 300 . The values of the response variable Y , denoted by $y_1, \dots, y_n$, are generated from the ESC distribution using the `qf` (6.10) and, for each value of n , all results are obtained from 2,000 Monte Carlo replications. Here, we

Figure 6.2. Density of the random errors Z generated for scenarios (a) 1, (b) 2 and (c) 3.

present and compare the results fitting the semiparametric ESC and normal models, for each scenario, where the model parameters are defined by

$$\text{ESC} \begin{cases} \mu_i = pb_{11}(x_{1i}, df) + \beta_{21} x_{2i} + \beta_{31} x_{3i}, & \sigma_i = \beta_{02}, \\ \nu_i = \exp(\beta_{03}) & \text{and} \quad \tau_i = \exp(\beta_{04}); \end{cases}$$

$$\text{Normal} \begin{cases} \mu_i = pb_{11}(x_{1i}, df) + \beta_{21} x_{2i} + \beta_{31} x_{3i} & \text{and} \quad \sigma_i = \exp(\beta_{02}), \end{cases}$$

where $pb(x_{1i}, df)$ represents a smooth P-spline function with respective degrees of freedom df to model X_1 . The purpose of this study is to verify the accuracy of the parameters associated with the explanatory variables X_1 , X_2 and X_3 considering different behaviors of the random errors. As the coefficients of the smoothing terms are meaningless, we only compare the estimates of the parameters β_{21} and β_{31} for the ESC and normal distributions. The biases and mean squared errors (MSEs) are evaluated and the results are reported in Table 6.1.

Table 6.1. The biases and MSEs of the ESC and normal parametric and semiparametric regression models based on 2,000 simulations for each scenario and $n=50, 150$ and 300 .

Scenario	n	Semiparametric ESC			Semiparametric normal		
		Parameter	Bias	MSE	Parameter	Bias	MSE
1	50	β_{21}	0.169	18.174	β_{21}	0.122	24.859
		β_{31}	0.019	4.304	β_{31}	0.046	5.946
	150	β_{21}	0.039	1.171	β_{21}	0.047	6.815
		β_{31}	0.015	0.301	β_{31}	0.014	1.736
	300	β_{21}	0.014	0.507	β_{21}	0.029	3.399
		β_{31}	0.002	0.120	β_{31}	0.013	0.940
2	50	β_{21}	0.038	1.671	β_{21}	0.039	2.023
		β_{31}	0.017	0.450	β_{31}	0.018	0.516
	150	β_{21}	0.025	0.340	β_{21}	0.009	0.639
		β_{31}	0.002	0.089	β_{31}	0.006	0.162
	300	β_{21}	0.007	0.145	β_{21}	0.010	0.288
		β_{31}	0.004	0.038	β_{31}	0.007	0.076
3	50	β_{21}	0.014	8.133	β_{21}	0.055	8.325
		β_{31}	0.065	2.020	β_{31}	0.055	2.103
	150	β_{21}	0.013	1.871	β_{21}	0.017	2.470
		β_{31}	0.012	0.482	β_{31}	0.014	0.605
	300	β_{21}	0.010	0.770	β_{21}	0.014	1.154
		β_{31}	0.004	0.208	β_{31}	0.000	0.284

The figures in Table 6.1 indicate that the MSEs of the MLEs of the parameters decay toward zero when the sample size n increases for both models and scenarios, as expected under first-order asymptotic theory. However, the MSEs of the semiparametric ESC model are smaller than those of the semiparametric normal model, thus indicating higher accuracy of the estimates of the parameters in the presence of bimodal and asymmetric random errors. Figure 6.3 display the fitted and generated terms

for the smooth functions under Scenarios 1, 2 and 3. We can note the inaccuracy of the estimates in the normal model due to the fact that this model is not suitable to fit bimodality and positive by and negative by skewed errors, respectively. Finally, we can conclude that the estimates of smoothing functions are affected when random errors are not properly estimated by the proposed models.

Figure 6.3. For scenarios (1) bimodal symmetric, (2) unimodal with positive skewness and (3) unimodal with negative skewness, the fitted and generated terms for the smooth functions based on 2,000 simulations of $n = 300$ for the (a) ESC and (b) normal models.

6.4 Applications

In this section, we provide two applications to real data to illustrate the flexibility of the semiparametric ESC regression models. The computations are performed using the `gamlss` subroutine in the R software. The scripts are available by the first author at <https://goo.gl/hAIcBF>. For both applications, the results are compared with those from the normal regression models.

6.4.1 Application: Body mass data

Consider the data of the Dutch growth study, a cross-sectional study that measures growth and development of the Dutch population between the ages 0 and 21 years for the regions North, East, West, South and City. The main objective of this study is to verify the relationship of the body mass index (T)

and the explanatory variable age (X_1). The full sample contains the measures of 7482 males and has in total 212 missing values for the explanatory variables, which are removed. To reduce the computational time of this analysis (approximately 75 hours for the full sample), we consider only the observations of the North totaling a sample of $n = 917$. For more details see Fredriks *et al.* (2000a) and Fredriks *et al.* (2000b).

We start the analysis considering only the response variable $Y = \log(T)$ by fitting the ESC and normal models. Table 6.2 lists the MLEs and the corresponding standard errors (SEs) in parentheses of the model parameters and the values of the statistics GD, AIC and BIC for the fitted models. Figure 6.4(a) provides the plots of the histogram of the current data and the fitted densities of the ESC and normal models. Clearly, the ESC model provides a good fit to these data.

Table 6.2. MLEs of the model parameters for the body mass data, the corresponding SEs (given in parentheses) and the GD, AIC and BIC statistics.

Model	Estimates				GD	AIC	BIC
ESC(μ, σ, ν, τ)	2.738 (0.022)	0.126 (0.055)	0.981 (0.215)	2.862 (2.761)	-865.5	-857.5	-838.2
Normal(μ, σ)	2.894 (0.005)	0.153 (0.027)			-829.7	-825.7	-816.0

Before fitting the regression models, as the preliminary analysis, we note in Figure 6.4(b) that the explanatory variable age has a nonlinear relationship with the response variable body mass index, indicating the use of nonlinear models. Further, we can also note that the variability of body mass index depends on age, thus indicating that the heteroscedastic models should be used to fit these data.

Figure 6.4. For the body mass data: (a) Empirical and estimated density for the ESC and normal models; (b) Observed y against age with fitted smooth curves.

Next, we present results of the semiparametric ESC and normal models using the steps proposed in Section 6.2.4 to select the additive terms. The model parameters are defined by

$$\begin{aligned} \text{ESC} \quad & \begin{cases} \mu_i = \beta_{01} + pb_{11}(X_{1i}, df), & \sigma_i = \exp(\beta_{02} + \beta_{12}x_{1i}), \\ \nu_i = \exp(\beta_{03}) \quad \text{and} \quad \tau_i = \exp(\beta_{04} + \beta_{14}x_{1i}); \end{cases} \\ \text{Normal} \quad & \begin{cases} \mu_i = \beta_{01} + pb_{11}(X_{1i}, df) \quad \text{and} \quad \sigma_i = \exp(\beta_{02} + \beta_{12}x_{1i}). \end{cases} \end{aligned}$$

Table 6.3 gives the MLEs, their approximate SEs and p -values obtained from the fitted ESC and normal semiparametric regression models to the body mass data. The coefficients of the smoothing terms have been omitted because they are meaningless.

Table 6.3. MLEs of the parameters and the approximate SEs from the fitted semiparametric ESC and normal models to the body mass data.

Semiparametric ESC				Semiparametric normal			
Parameter	Estimate	SE	<i>p</i> -value	Parameter	Estimate	SE	<i>p</i> -value
β_{01}	2.745	0.017	<0.001	β_{01}	2.738	0.005	<0.001
$pb_{11}(x_{1i}, 10.35)$				$pb_{11}(X_{1i}, 9.55)$			
β_{02}	-3.002	0.089	<0.001	β_{02}	-2.433	0.043	<0.001
β_{12}	0.030	0.006	<0.001	β_{12}	0.015	0.003	<0.001
β_{03}	-0.234	0.088	0.008	-	-	-	-
β_{04}	-0.049	0.210	0.813	-	-	-	-
β_{14}	0.053	0.021	0.013	-	-	-	-
$GD = -1606.0$ $AIC = -1575.3$ $BIC = -1501.3$				$GD = -1559.9$ $AIC = -1536.8$ $BIC = -1481.0$			

The results presented in Table 6.3 reveal that the semiparametric ESC model has lower GD, AIC and BIC statistics compared to the semiparametric normal model. To check the adequacy of the fitted distributions given in Table 6.3, we present in Figure 6.5 the worm plots considering four ranges of X_1 and, to compare the assumptions of the models, we also provide the index plots for the quantile residuals. Figure 6.5(b) indicates failure for modelling the kurtosis and skewness for the normal model. We may note in Figures 6.5(c)-(d) that the quantile residuals follow approximately a normal distribution but the semiparametric normal model has most points out of the range $[-3, 3]$, thus indicating the flexibility of the ESC model.

The partial effects of X_1 in the parameters of the fitted semiparametric ESC regression model are presented in Figure 6.6, which appear to be consistent with the effects presented in Figure 6.4(b). Figure 6.6(a) indicates that the log of the body mass index increases quickly until one year age, and then decreases at a slower rate until 6 years age, and after that increases until 21 years age. The plots in Figure 6.6(b)-(c) reveal that the variability and skewness of Y increases when x_1 increases.

Next, we compute the case deletion $LD_i(\theta)$ measure for the body mass data. Figures 6.7(a)-(b) reveals the influence measure index plots and the values of Y against X_1 with some possible influential points highlighted, respectively. From these plots, we can note that the cases 263, 447 and 442 appear as possible influential observations. Note that the cases 263 and 447 are also detected in the quantile residual plots (see Figure 6.5(c)). In fact, the case 263 has the lowest value of Y and the case 447 has a highest value of Y for the range $10 < X_1 < 15$.

Finally, Figure 6.8(a) displays the fitted semiparametric ESC regression model to the body mass data with some fitted conditional densities for different values of X_1 . We can note in this plot that the fitted ESC semiparametric regression model has unimodal shapes with null and positive skewness, e.g. for $x_{1i} = 7$ and $x_{1i} = 19$, respectively. Figure 6.8(b) provides five fitted percentile curves $u \times (10, 25, 50, 75, 90, 95)$ for Y against the eruption waiting time. We conclude the semiparametric ESC regression can be chosen as the best model.

6.5 Eruption data

In this section, we provide an analysis of the data on the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. The data consist of $n = 272$ observations on waiting times between eruptions and the duration of the eruption. Let the response variable t_i be the i th recorded duration of eruption and the explanatory variable x_{i1} the waiting time for the eruption. This data set can be obtained using `data(faithful)` in the R software. We note that there are many versions of these data: Azzalini and Bowman (1990) used a more complete version.

We consider the random variable $Y = \log(T)$ having the ESC and normal distributions. Table 6.4 gives the MLEs (and the corresponding SEs in parentheses) of the model parameters and the values

Figure 6.5. To the body data: The worm plot for the semiparametric (a) ESC and (b) normal models and the index plot of the quantile residuals for the semiparametric (c) ESC and (d) normal models.

Figure 6.6. The fitted terms (a) μ , (b) σ and (c) τ for the semiparametric ESC regression model given in Table 6.3.

of the statistics GD, AIC and BIC for the fitted models. Figure 6.9(a) provides the plots of the histogram of the current data and the fitted densities of the ESC and normal models. Table 6.4 and Figure 6.9(a) indicate that the ESC model provides a good fit to these data.

To propose the regression models, as a preliminary analysis, we present in Figure 6.9(b) the values of Y against X_1 , for which we note that the response variable Y has a nonlinear relationship with

Figure 6.7. For body mass data: (a) Index plots for $|LD_i(\theta)|$ and (b) Observed Y against X_1 .

Figure 6.8. For the semiparametric ESC regression model fitted to the eruption data: (a) smoothed scatterplot diagram showing how the fitted conditional distribution of the response variable Y changes for different values of X_1 ; (b) fitted percentile curves for $u \times 100 = (5, 25, 50, 75, 95)$ against X_1 .

Table 6.4. MLEs of the model parameters for the eruption data, the corresponding SEs (given in parentheses) and the GD, AIC and BIC statistics.

Model	Estimates				GD	AIC	BIC
ESC(μ, σ, ν, τ)	1.044 (0.008)	0.076 (0.059)	0.010 (0.306)	1.545 (0.361)	-88.1	-80.1	-65.7
Normal(μ, σ)	1.185 (0.022)	0.374 (0.062)			237.0	241.0	248.3

the explanatory variable X_1 , so that the nonlinear models are required.

Using the steps proposed in Section 6.2.4 to select additive terms, we present and compare the results of the semiparametric ESC and normal models, where the model parameters are defined by

$$\mathbf{ESC} \begin{cases} \mu_i = \beta_{01} + pb_{11}(X_{1i}, df), & \sigma_i = \exp[\beta_{02} + pb_{12}(X_{1i}, df)], \\ \nu_i = \exp(\beta_{03} + \beta_{13}x_{1i}) & \text{and } \tau_i = \exp(\beta_{04} + \beta_{14}x_{1i}); \end{cases}$$

$$\mathbf{Normal} \begin{cases} \mu_i = \beta_{01} + pb_{11}(X_{1i}, df) & \text{and } \sigma_i = \exp[\beta_{02} + pb_{12}(X_{1i}, df)]. \end{cases}$$

Table 6.5 provides the MLEs, their approximate SEs and p -values obtained from the fitted ESC and normal semiparametric regression models. The coefficients of the smoothing terms have been omitted to

Figure 6.9. For the eruption data: (a) The empirical and the estimated densities for the ESC and normal models; (b) Observed Y against X_1 with smooth fitted curves.

avoid erroneous interpretations.

Table 6.5. MLEs of the model parameters and the corresponding SEs from the fitted semiparametric ESC and normal regression models to the eruption data.

Semiparametric ESC				Semiparametric normal			
Parameter	Estimate	SE	p -value	Parameter	Estimate	SE	p -value
β_{01}	1.798	0.017	<0.001	β_{01}	-0.580	0.031	<0.001
$pb_{11}(x_{1i}, 11.21)$				$pb_{11}(X_{1i}, 9.15)$			
β_{02}	0.967	0.334	0.004	β_{02}	-1.584	0.227	<0.001
$pb_{12}(x_{1i}, 5.47)$				$pb_{12}(X_{1i}, 5.67)$			
β_{03}	3.293	0.517	<0.001	-	-	-	-
β_{13}	-0.049	0.007	<0.001	-	-	-	-
β_{04}	5.136	0.340	<0.001	-	-	-	-
β_{14}	-0.080	0.004	<0.001	-	-	-	-
$GD = -507$ $AIC = -465$ $BIC = -391$				$GD = -462$ $AIC = -432$ $BIC = -379$			

To verify the adequacy and the assumptions of the proposed models in Table 6.5, we present in Figure 6.10 the worm plots for four ranges of X_1 and the index plots for the quantile residuals. Figure 6.10(a)-(b) indicates a good fit of the ESC model for all ranges of X_1 and failure for modelling the skewness for the normal model. We can note in Figures 6.10(c)-(d) that the quantile residuals follow approximately a normal distribution and the semiparametric ESC model does not have points out of the range $[-3, 3]$, thus indicating the flexibility of the new semiparametric model.

The partial effects of X_1 in the parameters of the semiparametric ESC regression model are presented in Figure 6.11. Figure 6.11(a) indicates that y_i decreases at a slower rate until $x_{1i} = 48$, then increases slowly until $x_{1i} = 59$, and after that increases quickly until $x_{1i} = 72$ and after this point increases again slowly. In addition, Figure 6.11(b) reveals that the variability of the log of eruption times decays rapidly for $x_1 > 60$. Figures 6.11(c)-(d) indicate that the distribution of Y has bimodality and negative skewness for high values of x_1 .

Next, we compute the case deletion measures $LD_i(\theta)$ for the eruption data. The results of such influence measure index plots are displayed in Figure 6.12(a). In Figure 6.12(b), we present the values of Y against X_1 and the points detected in the influential analysis. From these plots, we note that the cases 19, 149, 211 and 265 are possible influential observations. Although these points have been detected in the influence analysis, the same does not appear as outlying observations in Figure 6.10, indicating again the flexibility of the new model.

Finally, Figure 6.13(a) reveals the semiparametric ESC regression model fitted to the eruption

Figure 6.10. To the eruption data: The worm plots for (a) semiparametric ESC and (b) semiparametric normal models and the index plots of the quantile residuals for (c) semiparametric ESC and (d) semiparametric normal models.

Figure 6.11. The fitted terms for (a) μ , (b) σ , (c) ν and (d) τ for the semiparametric ESC regression model.

Figure 6.12. For eruption data: (a) Index plots for $|LD_i(\theta)|$ and (b) Observed Y against X_1 .

data on a smoothed scatterplot diagram. We can note in this plot that the fitted semiparametric ESC regression model takes different shapes for different values of X_1 as bimodal and unimodal with positive skewness. Figure 6.13(b) displays five fitted percentile curves $u \times (10, 25, 50, 75, 90, 95)$ for the logarithms of recorded duration of eruptions against waiting times for the eruption. We can conclude that the semiparametric ESC could be chosen as the best model to the current data.

Figure 6.13. For the semiparametric ESC regression model fitted to the eruption data: (a) smoothed scatterplot diagram showing how the fitted conditional distribution of the response variable Y changes for different values of X_1 ; (b) fitted percentile curves for $u \times 100 = (5, 25, 50, 75, 95)$ against X_1 .

6.6 Conclusions

The semiparametric ESC 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 skewness and they can each be modelled as parametric or smooth nonparametric functions of explanatory variables. Procedures for fitting the semiparametric ESC regression model and for model diagnostics are included in the GAMLSS package and available from the authors. Two real data sets are used to illustrate the importance of the semiparametric ESC regression model, showing that it provides better performance than the usual methods in the presence of bimodal and asymmetric random errors.

References

- Atkinson, A.C. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford: Clarendon Press.
- Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics*, **39**, 357–365.
- 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., Lachos, V.H. and Ortega, E.M. (2010). A nonlinear regression model with skew-normal errors. *Statistical Papers*, **51**, 547–558.
- Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society B*, **48**, 133–169.
- Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.
- Cooray, K. (2013). Exponentiated sinh Cauchy distribution with applications. *Communications in Statistics-Theory and Methods*, **42**, 3838–3852.
- Cordeiro, G.M., Ortega, E.M.M. and Ramires, T.G. (2015). A new generalized Weibull family of distributions: mathematical properties and applications. *Journal of Statistical Distributions and Applications*, **2**, 1–25.
- Cysneiros, F.J.A., Cordeiro, G.M. and Cysneiros, A.H.M.A. (2010). Corrected maximum likelihood estimators in heteroscedastic symmetric nonlinear models. *Journal of Statistical Computation and Simulation*, **80**, 451–461.
- Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *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.
- Fredriks, A.M., Van Buuren, S., Burgmeijer, R.J., Meulmeester, J.F., Beuker, R.J., Brugman, E. and Wit, J.M. (2000a). Continuing positive secular growth change in The Netherlands 1955 - 1997. *Pediatric research*, **47**, 316–323.
- Fredriks, A.M., van Buuren, S., Wit, J.M. and Verloove-Vanhorick, S.P. (2000b). Body index measurements in 1996- 7 compared with 1980. *Archives of Disease in Childhood*, **82**, 107–112.
- Lachos, V.H., Bandyopadhyay, D. and Garay, A.M. (2011). Heteroscedastic nonlinear regression models based on scale mixtures of skew-normal distributions. *Statistics & Probability Letters*, **81**, 1208–1217.
- Lee, Y., Nelder, J.A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. CRC Press.
- Nakamura, L.R, Rigby, R.A, Stasinopoulos, D.M, Leandro, R.A and Villegas, C. (2016) A new extension of the Birnbaum-Saunders distribution using the GAMLSS framework. *Statistical Modelling*. (submitted)
- Ortega, E.M., Cordeiro, G.M. and Kattan, M.W. (2013). The log-beta Weibull regression model with application to predict recurrence of prostate cancer. *Statistical Papers*, **54**, 113–132.

- Ortega, E.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 Hamedani, G. (2013). The beta generalized half-normal geometric distribution. *Studia Scientiarum Mathematicarum Hungarica*, **50**, 523–554.
- 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.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. CRC Press.
- 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.
- Vanegas, L.H. and Paula, G.A. (2015). A semiparametric approach for joint modeling of median and skewness. *TEST*, **24**, 110–135.
- Vanegas, L.H. and Paula, G.A. (2015). An extension of log-symmetric regression models: R codes and applications. *Journal of Statistical Computation and Simulation*, **86**, 1709–1735.
- 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.
- Xu, D., Zhang, Z. and Du, J. (2015). Skew-normal semiparametric varying coefficient model and score test. *Journal of Statistical Computation and Simulation*, **85**, 216–234.

7 ESTIMATING NONLINEAR EFFECTS IN REGRESSION MODELS WITH LONG-TERM SURVIVORS

Abstract: Nonlinear effects between explanatory and response variables are increasingly present in new surveys. In this paper, we propose a flexible four-parameter 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. Bias caused by non incorporating of such non-linear effects in the model are investigated using Monte Carlo simulations. We discuss diagnostic measures and methods to select additive terms and its 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; P-spline; residual analysis; semi-parametric models.

7.1 Introduction

The objective of this study is to analyze censored data with the presence of long-duration individuals in which explanatory variables have nonparametric behavior in relation to the failure time. 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, some explanatory variables can present nonlinear behavior, i.e., behavior that does not have defined or known form. 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 test hypothesis 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 7.1(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 7.6, in which a thorough study is conducted. Note that the the proportion of cured individuals increases and then decreases, as 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 7.1(b), where we display the fitted cure rate proportions for each category of *age* using nonparametric techniques.

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 and it will depend on the priori given by the researcher, which is not always possible. 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

Figure 7.1. (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.

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 model framework is called in literature as the *generalized additive model for location, scale and shape* (GAMLSS) (Rigby and Stasinopoulos, 2005). We also consider, for each model parameter, smoothing techniques to capture nonlinear effects existent in the continuous explanatory variables.

We consider that the failure times follow the log-sinh Cauchy (LSC) distribution (Ramires *et al.*, 2016) and propose a new model called the *log-sinh Cauchy cure rate* (LSCcr) model. The paper is organized as follows. In Section 7.2, 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 7.4. In Section 7.5, we discuss methods for generating random values and Monte Carlo simulations on the finite sample behavior of the maximum likelihood estimates (MLEs). An application to breast cancer data presented in Section 7.6 illustrates the flexibility of the proposed semi-parametric regression model. Computational implementation and instructions for fitting the proposed model are given in the Appendix. Finally, we offer some conclusions in Section 7.7.

7.2 The Log sinh Cauchy GAMLSS 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 paper by Chen *et al.* (1999), Tsodikov *et al.* (2003) and the article by Cooner *et al.* (2007) could be mentioned as key references. Recently, other works dealt with cure rate models. For example, 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) proposed a new long-term survival model with interval-censored data, Cordeiro *et al.* (2016) proposed the negative binomial Birnbaum-Saunders model with

long-term survivors, and Ortega *et al.* (2015) defined a power series beta Weibull regression model for predicting breast carcinoma.

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 $i = 1, \dots, n$) be the indicator denoting that the i th individual is susceptible ($N_i = 1$) or non-susceptible ($N_i = 0$), 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

$$S_{pop}(t_i) = \tau + (1 - \tau)S(t_i|N_i = 1), \quad (7.1)$$

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 (7.1) is given by

$$f_{pop}(t_i) = -\frac{dS_{pop}(t_i)}{dt} = (1 - \tau)f(t_i|N_i = 1), \quad (7.2)$$

where $f(t_i|N_i = 1)$ is the baseline pdf for the susceptible individuals. Equations (7.1) and (7.2) are improper functions, since $S_{pop}(t)$ is not a proper survival function. We can omit sometimes 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

$$f(t; \mu, \sigma, \nu) = \frac{\nu}{t \sigma \pi} \frac{\cosh\left(\frac{\log(t) - \mu}{\sigma}\right)}{\left[\nu^2 \sinh^2\left(\frac{\log(t) - \mu}{\sigma}\right) + 1\right]}, \quad (7.3)$$

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 (7.3) is given by

$$S(t; \mu, \sigma, \nu) = 1 - \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh \left(\frac{\log(t) - \mu}{\sigma} \right) \right] \right\}. \quad (7.4)$$

7.2.1 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 (7.1) is given by (7.4), which gives

$$S_{pop}(t; \mu, \sigma, \nu, \tau) = 1 + (\tau - 1) \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[\nu \sinh(w) \right] \right\}, \quad (7.5)$$

where $w = \frac{\log(t) - \mu}{\sigma}$. We can omit sometimes the dependence on the parameters as, for example, $S_{pop}(t) = S_{pop}(t; \mu, \sigma, \nu, \tau)$. The pdf corresponding to (7.5) is given by

$$f_{pop}(t) = \frac{(1 - \tau) \nu}{\sigma \pi t} \frac{\cosh(w)}{[\nu^2 \sinh^2(w) + 1]}. \quad (7.6)$$

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 (7.6) 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 7.2 and 7.3, respectively.

Figure 7.2. 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 7.3. 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 7.2(a)-(b) reveal clearly the symmetric and bi-modality effects due to the parameters σ and τ , respectively, and different effects of the cured probability τ . Further, Figures 7.3(a)-(b) indicate that the hrf of T can have decreasing, unimodal and bimodal shapes. We can note in Figure 7.3(b) that the values of the hrf are smaller in the presence of the proportion of cured individuals but still assuming the same characteristics.

7.3 The LSCcr GAMLSS

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 Stasinopoulou, 2005) allows all parameters of the conditional distribution of t be modelled as parametric 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 misuse 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 $T \sim LSCr(y; \theta)$, where $\theta = (\mu, \sigma, \nu, \tau)^T$ denotes the vector of parameters of the pdf (7.6). Consider independent observations t_i conditional on the parameter vector θ_i (for $i = 1, \dots, n$) having pdf $f(t_i; \theta_i)$, where $\theta^T = (\mu^T, \sigma^T, \nu^T, \tau^T)$ is a vector of parameters related to the response variable. The GAMLSS allows the user to model all parameters in θ 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 θ using appropriate link functions as

$$\theta = \begin{bmatrix} \mu \\ \sigma \\ \nu \\ \tau \end{bmatrix} = \begin{bmatrix} g_1 \left(\mathbf{X}_1 \beta_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1}) \right) \\ g_2 \left(\mathbf{X}_2 \beta_2 + \sum_{j=1}^{J_2} h_{j2}(\mathbf{x}_{j2}) \right) \\ g_3 \left(\mathbf{X}_3 \beta_3 + \sum_{j=1}^{J_3} h_{j3}(\mathbf{x}_{j3}) \right) \\ g_4 \left(\mathbf{X}_4 \beta_4 + \sum_{j=1}^{J_4} h_{j4}(\mathbf{x}_{j4}) \right) \end{bmatrix}, \quad (7.7)$$

where $g_k(\cdot)$ for $k = 1, 2, 3, 4$ denote the injective and twice continuously differentiable monotonic link functions, $\beta_k = (\beta_{0k}, \beta_{1k}, \dots, \beta_{m_k k})^T$ is a parameter vector of length $(m_k + 1)$, m_k denotes the number of explanatory variables related to the k th 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 Stasinopoulou (2005) proved that each smoothing function $h_{jk}(\cdot)$ can be expressed as a random effects model, i.e., $h_{jk}(\cdot) = \mathbf{Z}_{jk} \gamma_{jk}$, where $\mathbf{Z}_{jk}$ is an $n \times q_{jk}$ matrix representing the B-spline basis design matrix and γ_{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).

7.4 Model selection

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

7.4.1 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 Stasinopoulos (2005) and Stasinopoulos and Rigby (2007) and available in the documentation of the GAMLSS package.

Consider a sample of n -independent observations $t_1, \dots, t_n$, 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 (7.7), we consider fixed the smoothing parameters λ_{jk} , and the fixed and random effects β and γ , respectively, are estimated by maximizing the penalized log-likelihood function

$$l_p = l(\theta) - \frac{1}{2} \sum_{k=1}^4 \sum_{j=1}^{J_k} \lambda_{jk} \gamma_{jk}^T \mathbf{P}_{jk} \gamma_{jk}, \quad (7.8)$$

where $\mathbf{P}_{jk}$ is a symmetric matrix that may depend on a vector of smoothing parameters (Rigby and Stasinopoulos, 2005). The non-penalized log-likelihood function $l(\theta) = \sum_{i \in F} \log f(t_i; \theta_i) + \sum_{i \in C} \log S(t_i; \theta_i)$ is given by

$$\begin{aligned} l(\theta) = & \sum_{i \in F} \left\{ \log(1 - \tau_i) + \log(\nu_i) - \log(\sigma_i \pi) - \log(t_i) + \log \cosh(w_i) - \log[1 + \nu_i^2 \sinh^2(w_i)] \right\} \\ & + \sum_{i \in C} \log \left(1 + (p_i - 1) \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu_i \sinh(w_i)] \right\} \right), \end{aligned} \quad (7.9)$$

where $w_i = [\log(t_i) - \mu_i]/\sigma_i$. The parameter vector $\theta = (\beta_1^T, \dots, \beta_4^T)^T$ is used to define the regression structures in (7.7) by specifying appropriate link functions for $g_k(\cdot)$, e.g., using the logit link function for $g_4(\tau)$, the parameter τ is related to the covariates by $\tau_i = \exp(\mathbf{X}_4[i,] \beta_4) / [1 + \exp(\mathbf{X}_4[i,] \beta_4)]$, where $\mathbf{X}_k[i,]$ denotes the i -th row of the model matrix $\mathbf{X}_k$. The fit of the LSCr model gives the vector of estimated cured proportion

$$\hat{\tau} = \frac{\exp[\mathbf{X}_4 \hat{\beta}_4 + \sum_{j=1}^{J_4} \hat{h}_{j4}(\mathbf{x}_{j4})]}{1 + \exp[\mathbf{X}_4 \hat{\beta}_4 + \sum_{j=1}^{J_4} \hat{h}_{j4}(\mathbf{x}_{j4})]}, \quad (7.10)$$

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

For each smoothing term selected, and any of the parameters of the LSCr distribution, there is one smoothing parameter λ associated with it. The smoothing parameters can be fixed or estimated from the data. We adopt the 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 Stasinopoulos, 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.

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 `edfA11(.)` function in the R software.

7.4.2 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 $GD = -2l(\hat{\theta})$, where $l_p(\hat{\theta})$ is the total log-likelihood function and the AIC and BIC criterion are obtained by $AIC = GD + 2df$ 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, $\hat{u}_i = 1 - S(t_i|\hat{\theta}_i)$ and $S(t_i|\hat{\theta}_i)$ is the survival function (7.5). For censored observations, considering a right censored continuous response, $\hat{u}$ is defined as a random value from a uniform distribution on the interval $[1 - S(t_i|\hat{\theta}_i), 1]$.

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 fit adequately 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.

7.4.3 Additive terms selection

For the LSCcr GAMLSS, the selection of the terms for all the parameters is performed using the stepwise 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 to select the additive terms for the τ parameter 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 , always using as fixed 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 μ , σ , ν and τ .

7.5 Simulation study

Consider the random variable T having pdf (7.3). By inverting $F(t) = 1 - S(t) = u$ in (7.4), we obtain the qf of the LSC distribution as

$$t = Q(u) = \exp \left(\mu + \sigma \operatorname{arcsinh} \left\{ \frac{1}{\nu} \tan [\pi (u - 0.5)] \right\} \right). \quad (7.11)$$

Equation (7.11) can be used for simulating $T \sim \text{LSC}(\mu, \sigma, \nu)$ 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 setting the parameters using the semi-parametric (7.7) structure.

We conduct a Monte Carlo simulation study to assess the finite sample behavior of the MLEs of the model parameters. We consider model (7.7), where the cure rate parameter τ has a nonlinear relationship with the explanatory variable X_1 . The total sample sizes are taken as $n = 200$ 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 has an effect in the parabola form in τ . For each level of X_1 , it was generated a sample size of length 20. The fixed values of τ , for each value of X_1 , are given in Table 7.1.

Table 7.1. Fixed values of the τ parameter of each level of the X_1 explanatory variable.

ν	0.2	0.35	0.4	0.55	0.6	0.6	0.55	0.4	0.35	0.2
X_1	1	2	3	4	5	6	7	8	9	10

The failure times T , denoted by $t_1, \dots, t_n$, are generated from the LSC distribution using the qf (7.11) and the censoring times C are randomly generated from the uniform distribution $C \sim [\max(T), 2 \operatorname{sd}(T)]$, where $\operatorname{sd}(T)$ denotes the standard deviation for 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, namely

- **Parametric** $\text{LSCcr}(\mu, \sigma, \nu, \operatorname{logit}[\beta_{04} + \beta_{14} X_1])$,
- **Semi-parametric** $\text{LSCcr}(\mu, \sigma, \nu, \operatorname{logit}[pb(X_1, df)])$,

where $pb(X_1, df)$ denotes a smooth P-spline function with corresponding degrees of freedom df to model X_1 . 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 are reported in Table 7.2. As the coefficients of the smoothing terms $pb(X_1, df)$ are meaningless, we only present average of the estimated degrees of freedom in this table.

Table 7.2. The AEs, biases and MSEs for the parametric and semi-parametric LSCcr regression models based on 1,000 simulations.

Parametric				Semi-parametric			
Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
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	-	-				

The figures in Table 7.2 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 model is not close to two, thus indicating that we have a nonlinear effect of X_1 in the cure rate parameter. Finally, taking into account the parameter estimates relative to the cure rate parameter for the parametric model, we note that β_{14} is approximately zero, erroneously indicating that the explanatory variable X_1 has no effect in the cure rate proportions. The main conclusion of this simulation study is that, when the regression model is unspecified correctly, i.e., not allowing that nonlinear effects can be estimated, erroneous conclusions can be drawn about the explanatory variables.

Figure 7.4 displays the generated and fitted effects for the parametric and semi-parametric models. We also present in this figure the box-plots of the GD, AIC and BIC statistics obtained in 1,000 simulations for both models. We can note that the estimates of the cure rate parameter $\hat{\tau}$ are more suitable for the semi-parametric model. Further, we can conclude that the semi-parametric model presents the lowest values of the GD, AIC and BIC statistics, thus indicating to be the most appropriate model to the current data.

Figure 7.4. For the fitted LSCr parametric and semi-parametric models: (a) the fitted and generated effect of X_1 in the τ parameter; (b) the goodness-of-fit statistics.

7.6 Predicting the cure rate of breast cancer

A prognosis is the doctor's 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 German (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 7.5 displays the empirical survival functions and the corresponding p -values of log-rank test for the categorical variables. We may observe in these plots that only menopausal status did not present significative difference between the survival curves. We also present the frequency histogram of three explanatory variables, progesterone receptor, tumor size and age, in Figure 7.6. These plots reveal that the highest concentration of progesterone receptor is in the range $[0,600]$, the average of tumor size is 29.3 and the average of age is 53.

Figure 7.5. Empirical survival functions and log-rank test for (a) *htreat* (b) *menostat* and (c) *tumgrad*.

Next, using the steps described in Section 7.4.3 to select the additive terms for the different parameters, we present 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

$$\begin{aligned}
 \mu_i &= \beta_{01} + \beta_{11}age + \beta_{21}prm + \beta_{31}htreat_1, \\
 \sigma_i &= \exp(\beta_{02} + \beta_{12}tumgrad_2 + \beta_{22}tumgrad_3), \\
 \nu_i &= \exp(\beta_{03}) \\
 \tau_i &= \text{logistic}[\beta_{04} + \beta_{14}prm + \beta_{24}tumsize + \beta_{34}htreat_1 + \beta_{44}tumgrad_2 + \beta_{54}tumgrad_3 + pb(age)],
 \end{aligned} \tag{7.12}$$

Figure 7.6. Frequency histogram of explanatory variables (a) progesterone receptor, (b) tumor size and (c) age.

Weibuller model

$$\mu_i = \exp[\beta_{01} + \beta_{11}tumgrad_2 + \beta_{21}tumgrad_3 + pb(esm) + pb(age)]$$

$$\sigma_i = \exp(\beta_{02} + \beta_{12}tumgrad_2 + \beta_{22}tumgrad_3) + \beta_{32}tumsie),$$

$$\nu_i = \text{logistic}[\beta_{04} + \beta_{14}prm + \beta_{24}tumsie + \beta_{34}htreat_1 + \beta_{44}tumgrad_2 + \beta_{54}tumgrad_3 + pb(age)],$$

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 7.3 lists the values of the GD, AIC and BIC statistics for the fitted models. We can conclude from the figures in this table that the LSCcr model provides a better fit than the Weibuller model.

Table 7.3. The GD, AIC and BIC statistics and corresponding degrees of freedom for the fitted LSCcr and Weibuller models.

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

Table 7.4 provides the MLEs, SEs and p -values obtained from the fitted LSCcr GAMLSS. The coefficients of the smoothing terms have been omitted to avoid erroneous interpretations. 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 , $tumsie$, $htreat$, $tumgrad$ and age are significant to fit the cure rate parameter being that age has a nonlinear effect in it.

Table 7.4. 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 7.7. From the model for μ , we may note that the recurrence free survival time t_i increases according the age (Panel (a)) and the progesterone receptor (Panel (b)) increase and is greater for patients treated with hormonal treatment with tamoxifen (Panel (c)). Regarding the scale parameter σ , as we

can see in Figure 7.8(a), the variability of t_i increases as the gradient tumor grade increases. For the cure rate parameter τ , we may conclude from Figure 7.8(b)-(f) that the probability of cure increases as 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 7.7. Fitted terms for the location μ parameter: (a) age, (b) progesterone receptor and (c) hormonal treatment.

Figure 7.8. 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.

Based on equation (7.10), the estimated cured proportions can be determined using the results obtained in (7.4) as $\hat{\tau}_i = \text{logistic}[3.224 + 0.002prm_i - 0.046tumsize_i + 0.519htreat_{1i} - 1.319tumgrad_{2i} - 1.678tumgrad_{3i} + pb(age, 5.183)]$. In Figure 7.9, we present the estimated cured proportions for different levels of the explanatory variables as function of *age*. We may note in this plot 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 not received hormonal treatment with tamoxifen. Finally, the probability of cure increases as age increases in the range $[20, 45]$, decreases in the age range $[45, 60]$ and then stabilizes as age is greater than 60.

Figure 7.9. The estimated cured proportions for each level of *tumgrad* and *htreat* as function of *age* by taking: (a) $\min(prm) = 0$ and $tumsiz = 60$ and (b) $prm = 200$ and $tumsiz = 10$.

Figure 7.10 shows 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 age (see Figure 7.8(d)) in the hrf.

Figure 7.10. 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 $tumsiz = 60$ and (b) $prm = 200$ and $tumsiz = 10$.

Figure 7.11 displays some residual plots that will help to verify the adequacy and the assumptions of the chosen fitted model given in (7.12). We also present in this figure the residual plots for the Weibuller model. Panel (a) and (d) indicate that the normalized quantile residuals have an approximately normal distribution. Panel (e) shows that there are a few points off the line in low end of the range. Finally, the WP presented in Panel (c) indicates that there are no evidences of inadequacies on it, since all the residuals fall in “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.

7.7 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-

Figure 7.11. 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.

parametric LSCcr *generalized additive model for location, scale and shape* (GAMLSS) and for model diagnostics are included in the GAMLSS package and they can be obtained from the authors under request. 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.

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. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*; 2001; 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. *Journal of Computational and Graphical Statistics* 1996; 5: 236-244.
- Eilers, P.H. and Marx, B.D. Flexible smoothing with B-splines and penalties. *Statistical Science* 1996; 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 PL, Page DL, Weaver D, Thor AD, Allred DC, Clark GM, et al. Prognostic factors in breast cancer: College of American Pathologists consensus statement 1999. *Archives of pathology & laboratory medicine* 2000; 124: 966-978.
- Gospodarowicz, M.K., O'Sullivan, B. and Sobin, L.H. (Eds.). (2006). Prognostic factors in cancer (pp. 165-168). Frankfurt: Wiley-Liss.
- Green, P.J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.
- Hashimoto, E.M., Ortega, 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. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2005; 54: 507-554.
- Rigby, R.A. and Stasinopoulos, D.M. Automatic smoothing parameter selection in GAMLSS with an application to centile estimation. *Statistical Methods in Medical Research* 2014; 23: 318-332.
- Stasinopoulos, D.M. and Rigby, R. A. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software* 2007; 23: 1-46.
- Schumacher, M., Bastert, G., Bojar, H., Huebner, K., *et al.* 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, 1994.
- Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J. and Stasinopoulos, D. Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics* 2012; 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.

8 CONCLUSION

The paper proposes the exponentiated log-sinh Cauchy (ELSC) distribution that can be used as an alternative to mixture distributions in modeling bimodal data. Various mathematical properties of the ELSC distribution are investigated. We show that it can accommodate various shapes of the skewness, kurtosis and bi-modality.

Based on the ELSC distribution, we propose a general class of exponentiated sinh Cauchy (ESC) regression models, where the mean, dispersion, skewness and bimodal parameters vary across observations through regression structures. The former class of regression models is very suitable for modeling censored and uncensored lifetime data. The proposed model serves as an important extension to several existing regression models and could be a valuable addition to the literature. We use the GAMLSS script in the R package to obtain the maximum likelihood estimates and perform asymptotic tests for the model parameters based on the asymptotic distribution of the estimates. We offer some interesting insights, especially regarding model checking, and provide applications of influence diagnostics (global, local and total influence) in the proposed class of regression models with censored data.

In the context of cure rate models, we introduce the exponentiated log-sinh Cauchy cure rate (ELSCcr) model that can be used as an alternative to mixture distributions in modeling bimodal data with or without the presence of immune proportion of individuals. Three real data examples prove empirically that the ELSCcr distribution is very flexible, parsimonious, and a competitive model that deserves to be added to existing distributions in modeling bimodal data. We also presents the parametric log-sinh Cauchy promotion time generalized additive model for location, scale and shape (LSCp GAMLSS) to estimate breast carcinoma mortality, assuming that the number of competing causes that can influence the survival time follows a Poisson distribution.

Considering the presence of non-linear effects occurred by explanatory variables, we present the semiparametric ESC regression model, where the parameters of the model can be modelled as parametric or smooth nonparametric functions of explanatory variables. Two real data sets are used to illustrate the importance of the semiparametric ESC regression model, showing that it provides better performance than the usual methods in the presence of bimodal and asymmetric random errors.

Finally, the semi-parametric *log-sinh Cauchy cure rate* (LSCcr) regression model was proposed, where the cure rate parameter can also be modeled using parametric or smooth nonparametric functions of explanatory variables. 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.

APPENDICES

Appendix A: Score functions for Chapter 6

Appendix A: Score functions

Let $\mathbf{U}^T(\boldsymbol{\theta}) = \partial l_p / \partial \boldsymbol{\theta} = [\mathbf{U}_{\beta_1}, \mathbf{U}_{\gamma_{j1}}, \mathbf{U}_{\beta_2}, \mathbf{U}_{\gamma_{j2}}, \mathbf{U}_{\beta_3}, \mathbf{U}_{\gamma_{j3}}, \mathbf{U}_{\beta_4}, \mathbf{U}_{\gamma_{j4}}]$ be the score functions of the likelihood (6.9), γ_{rjk} the r th element of the q_{jk} -dimensional vector $\boldsymbol{\gamma}_{jk}$, $\beta_{l_k j}$, for $l_k = 0, 1, \dots, p_k$, the l th element of the vector $\boldsymbol{\beta}_j$ and $p_{jk}[r, s]$ the elements of the matrix P_{jk} . The elements of the $\mathbf{U}(\boldsymbol{\theta})$ are given by

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_{l_1 1}} = u_\mu(\beta_{l_1 1}) &= \sum_{i \in F} [\dot{g}_1^{-1}(\mu_i)]_{\beta_{l_1 1}} \left[\frac{\nu_i^2 \sinh(2w_i)}{\sigma_i K_i} - \frac{\tanh(w_i)}{\sigma_i} - (\tau_i - 1) \frac{\nu_i \cosh(w_i)}{\pi \sigma B_i K_i} \right] \\ &+ \sum_{i \in C} [\dot{g}_1^{-1}(\mu_i)]_{\beta_{l_1 1}} \frac{\tau_i \nu_i \cosh(w_i) B_i^{\tau_i - 1}}{\pi \sigma_i K_i (1 - B_i^{\tau_i})}, \end{aligned}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \gamma_{rj1}} = u_\mu(\gamma_{rj1}) - \sum_{j=1}^{J_1} \sum_{s=1}^{q_{j1}} \lambda_{j1} p_{j2}[r, s] \gamma_{rj1},$$

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_{l_2 2}} = u_\sigma(\beta_{l_2 2}) &= \sum_{i \in F} [\dot{g}_2^{-1}(\sigma_i)]_{\beta_{l_2 2}} \left[\frac{-1 - w_i \tanh(w_i)}{\sigma_i} + \frac{\nu_i^2 w_i}{\sigma_i K_i} \sinh(2w_i) - (\tau_i - 1) \frac{\nu_i w_i \cosh(w_i)}{\pi \sigma_i B_i K_i} \right] \\ &- \sum_{i \in C} [\dot{g}_2^{-1}(\sigma_i)]_{\beta_{l_2 2}} \frac{\tau_i \nu_i w_i B_i^{\tau_i - 1} \cosh(w_i)}{\pi \sigma_i K_i (1 - B_i^{\tau_i})}, \end{aligned}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \gamma_{rj2}} = u_\sigma(\gamma_{rj2}) - \sum_{j=1}^{J_2} \sum_{s=1}^{q_{j2}} \lambda_{j2} p_{j2}[r, s] \gamma_{rj2},$$

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_{l_3 3}} = u_\nu(\beta_{l_3 3}) &= \sum_{i \in F} [\dot{g}_3^{-1}(\nu_i)]_{\beta_{l_3 3}} \left[\frac{1}{\nu_i} - \frac{2\nu_i \sinh^2(w_i)}{K_i} + (\tau_i - 1) \frac{\sinh(w_i)}{\pi B_i K_i} \right] \\ &+ \sum_{i \in C} [\dot{g}_3^{-1}(\nu_i)]_{\beta_{l_3 3}} \frac{-\tau_i B_i^{\tau_i - 1} \sinh(w_i)}{\pi K_i (1 - B_i^{\tau_i})}, \end{aligned}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \gamma_{rj3}} = u_\nu(\gamma_{rj3}) - \sum_{j=1}^{J_3} \sum_{s=1}^{q_{j3}} \lambda_{j3} p_{j3}[r, s] \gamma_{rj3},$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_{l_4 4}} = u_\tau(\beta_{l_4 4}) = \sum_{i \in F} [\dot{g}_4^{-1}(\tau_i)]_{\beta_{l_4 4}} \left[\frac{1}{\tau_i} + \log(B_i) \right] + \sum_{i \in C} [\dot{g}_4^{-1}(\tau_i)]_{\beta_{l_4 4}} \frac{-B_i^{\tau_i}}{1 - B_i^{\tau_i}} \log(B_i) \quad \text{and}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \gamma_{rj4}} = u_\tau(\gamma_{rj4}) - \sum_{j=1}^{J_4} \sum_{s=1}^{q_{j4}} \lambda_{j4} p_{j4}[r, s] \gamma_{rj4},$$

where $[\dot{g}_k^{-1}(\cdot)]_{\psi_k} = \frac{\partial [g_k^{-1}(\cdot)]}{\partial \psi_k}$, for $k = 1, \dots, 4$, $B_i = \frac{1}{2} + \frac{1}{\pi} \arctan[\nu_i \sinh(w_i)]$, $K_i = \nu_i^2 \sinh^2(w_i) + 1$ and $w_i = [y_i - \mu_i] / \sigma_i$.

Appendix B: Computational codes for Chapter 7

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)
```