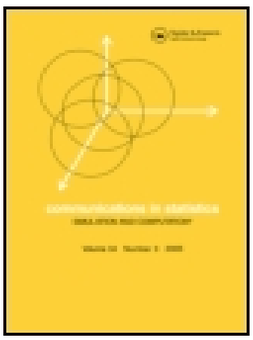
New regression model with four regression structures and computational aspects

Ramires, Thiago G.; Ortega, Edwin M. M.; Cordeiro, Gauss M.; Paula, Gilberto A. & HENS, Niel (2018) New regression model with four regression structures and computational aspects. In: COMMUNICATIONS IN STATISTICS-SIMULATION AND COMPUTATION, 47(7), p. 1940-1962.

DOI: 10.1080/03610918.2017.1332212
Handle: http://hdl.handle.net/1942/27427



Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

New regression model with four regression structures and computational aspects

Thiago G. Ramires, Edwin M.M. Ortega, Gauss M. Cordeiro, Gilberto A. Paula & Niel Hens

To cite this article: Thiago G. Ramires, Edwin M.M. Ortega, Gauss M. Cordeiro, Gilberto A. Paula & Niel Hens (2017): New regression model with four regression structures and computational aspects, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2017.1332212

To link to this article: http://dx.doi.org/10.1080/03610918.2017.1332212

Accepted author version posted online: 26 May 2017.

New regression model with four regression structures and computational aspects

Thiago G. Ramires
Department of Exact Sciences, University of São Paulo, Brazil
Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-Biostat), University of Hasselt, Belgium;

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

Gauss M. Cordeiro
Department of Statistics, Federal University of Pernambuco, Brazil

Gilberto A. Paula
Department of Statistics, Institute of Mathematics and Statistics, USP, Brazil

Niel Hens
Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-Biostat), University of Hasselt, Belgium;
Centre for Health Economic Research and Modelling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Belgium

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.

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 flexility 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 being 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

$$F(t) = G(t)^{\tau}, \tag{1}$$

where $G(t)$ represents the baseline cdf and $\alpha > 0$ denotes the shape parameter. By differentiating (1), the corresponding probability density function (pdf) becomes

$$f(x) = \tau G(t)^{\tau-1} g(t), \tag{2}$$

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 (2) by defining the four-parameter *exponentiated log-sinh Cauchy* (ELSC) distribution, whose pdf (for $t > 0$) is given by

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

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

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

The pdf and survival function corresponding to (4) are given by

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

and

$$S(y;\mu,\sigma,\nu,\tau) = \frac{(2\pi)^{\tau} - \left\{\pi + 2\arctan\left[\nu\,\sinh\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\tau}}{(2\pi)^{\tau}}, \tag{6}$$

respectively. The ESC distribution (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 (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 Appendix.

The sections are organized as follows. In Section 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. Section 4 contains some Monte Carlo simultaneously on the finite sample behavior of the MLEs. In Section 5, we assess the behavior of the MLEs of the parameters in the ESC regression model when it is poorly specified. In Section 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 7 to illustrate the flexibility of the proposed class of regression models for censored and uncensored data. Finally, Section 8 offers some conclusions.

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 $z \in \mathbb{R}$) reduces to

$$f(z;\nu,\tau) = \tau\, g_{\mathcal{SC}}(z)\, G_{\mathcal{SC}}(z)^{\tau-1} = \frac{\tau\nu}{\pi}\, \frac{\cosh(z)}{\nu^2\,\sinh^2(z)+1} \left\{\frac{1}{2} + \frac{1}{\pi}\arctan\left[\nu\,\sinh(z)\right]\right\}^{\tau-1}, \tag{7}$$

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

$$G_{\mathcal{SC}}(z) = \left\{\frac{1}{2} + \frac{1}{\pi}\arctan\left[\nu\,\sinh(z)\right]\right\} \qquad \text{and} \qquad g_{\mathcal{SC}}(z) = \frac{\nu}{\pi}\,\frac{\cosh(z)}{\nu^2\,\sinh^2(z)+1}, \tag{8}$$

respectively.

Plots of the density function (7) for selected parameter values are displayed in Figure 1. Equation (7) for the standardized ESC distribution will be used in Section 3.1 to specify the error distribution of the proposed regression model.

2.1 Expansion of the quantile function

Inverting $F(y) = u$ in (4) gives the qf of Y

$$Y = Q_Y(u) = \mu + \sigma\,\text{arcsinh}\left\{\frac{1}{\nu}\tan\left[\pi\left(u^{1/\tau} - 0.5\right)\right]\right\}. \tag{9}$$

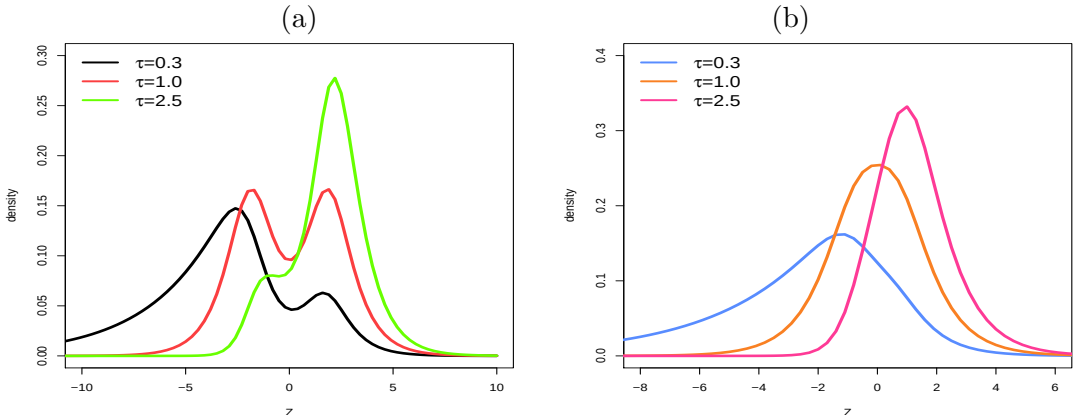


Figure 1: Plots of the density function (7) for some values of τ: (a) $\nu = 0.3$; (b) $\nu = 0.8$.

The qf $Q_Z(u)$ of Z, which has the standardized ESC density function (7), can be obtained from (9) with $\mu = 0$ and $\sigma = 1$. The qf of the standardized SC distribution, say $Q_{SC}(u)$, also follows (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 (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 (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 $b_0 = 1$, $b_1 = 2\nu^2 - 1$, $b_2 = 16\nu^4 - 20\nu^2 + 9$, $b_3 = 272\nu^6 - 616\nu^4 + 630\nu^2 - 225$, $b_4 = 7936\nu^8 - 28160\nu^6 + 48384\nu^4 - 37800\nu^2 + 11025, \ldots$

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

$$Q_Z(u) = \sum_{j=0}^{\infty} p_j \, u^{j/\tau}, \tag{10}$$

where the coefficients

$$p_j = \sum_{k=j}^{\infty} (-0.5)^{2k+1-j} \binom{2k+1}{j} c_k \tag{11}$$

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

2.2 Moments

Let $\mu'_s = E(Z^s)$ be the sth ordinary moment of Z with pdf (7). We have

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

Replacing $Q_{SC}(u)$ (eq. (10) when $\tau = 1$) in the last equation, we obtain

$$\mu'_n = \tau \int_0^1 \left(\sum_{j=0}^{\infty} p_j \, u^j \right)^s u^{\tau-1} du. \tag{12}$$

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

$$\left(\sum_{i=0}^{\infty} a_i \, u^i \right)^n = \sum_{i=0}^{\infty} b_{n,i} \, u^i, \tag{13}$$

where the coefficients $b_{n,i}$ (for $i = 1, 2, \ldots$) 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 $a_0, \ldots, a_i$.

Based on equation (13), equation (12) can be rewritten as

$$\mu'_n = \tau \sum_{j=0}^{\infty} e_{s,j} \int_0^1 u^{j+\tau-1} du = \sum_{j=0}^{\infty} \frac{\tau}{\tau + j} \, e_{s,j}, \tag{14}$$

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 (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 2 and 3 for selected values of τ as functions of ν and for selected values of ν as functions of τ, respectively.

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 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.

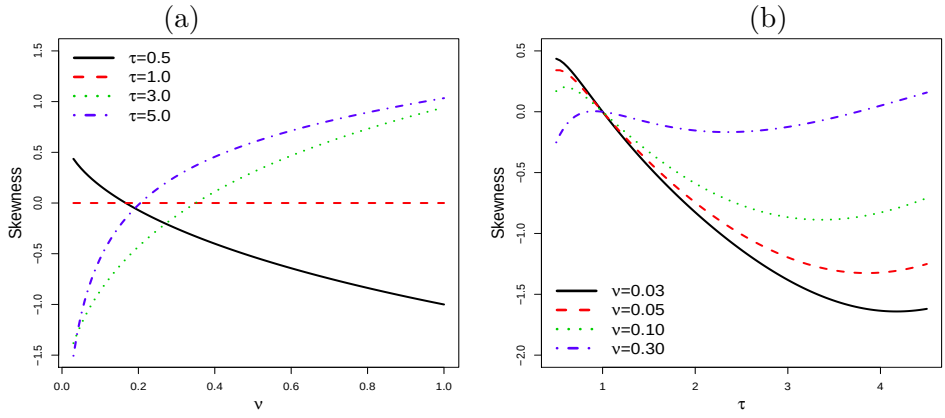


Figure 2: Skewness of the ESC distribution: (a) Function of ν for some values of τ. (b) Function of τ for some values of ν.

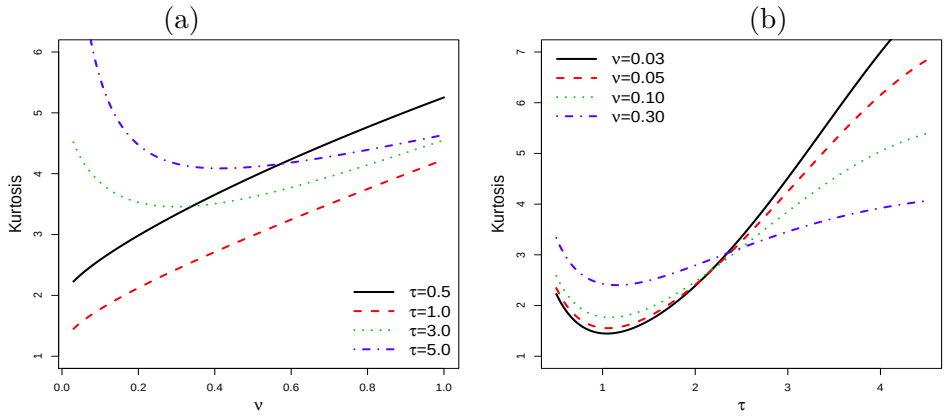


Figure 3: Kurtosis of the ESC distribution: (a) Function of ν for some values of τ. (b) Function of τ for some values of ν.

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 Stasinopouls, 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.

3.1 Definition

Let $\boldsymbol{\theta}^T = (\mu, \sigma, \nu, \tau)$ denote the vector of parameters of the pdf (5). We consider that independent observations y_i conditional on $\boldsymbol{\theta}_i$ (for $i = 1, \ldots, n$), with pdf $f(y_i; \boldsymbol{\theta}_i)$, where ${\boldsymbol{\theta}_i}^T = (\mu_i, \sigma_i, \nu_i, \tau_i)$ 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

$$y_i = \mu_i + \sigma_i\, z_i, \qquad i = 1, \ldots, n, \tag{15}$$

where the random error z_i follows the density function $f(z_i; \nu_i, \tau_i)$ given by (7) and $Z_i = (Y_i - \mu_i)/\sigma_i$. We define the parameter vector $\boldsymbol{\theta}$ using appropriate link functions as

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\sigma} \\ \boldsymbol{\nu} \\ \boldsymbol{\tau} \end{bmatrix} = \begin{bmatrix} g_1(\mathbf{X}_1\boldsymbol{\beta}_1) \\ g_2(\mathbf{X}_2\boldsymbol{\beta}_2) \\ g_3(\mathbf{X}_3\boldsymbol{\beta}_3) \\ g_4(\mathbf{X}_4\boldsymbol{\beta}_4) \end{bmatrix} \text{ or } \theta_i = \begin{bmatrix} \mu_i \\ \sigma_i \\ \nu_i \\ \tau_i \end{bmatrix} = \begin{bmatrix} g_1(\beta_{01} + x_1[i,2]\beta_{11} + \ldots + x_1[i,p_1+1]\beta_{p_1 1}) \\ g_2(\beta_{02} + x_2[i,2]\beta_{12} + \ldots + x_2[i,p_2+1]\beta_{p_2 2}) \\ g_3(\beta_{03} + x_3[i,2]\beta_{13} + \ldots + x_3[i,p_3+1]\beta_{p_3 3}) \\ g_4(\beta_{04} + x_4[i,2]\beta_{14} + \ldots + x_4[i,p_4+1]\beta_{p_4 4}) \end{bmatrix}, \tag{16}$$

where p_k denotes the number of explanatory variables related to the kth 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, $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{p_k k})^T$ 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 (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.

3.2 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 (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 (5) and $S(y_i; \boldsymbol{\theta}_i)$ is the survival function in (6). The log-likelihood function for $\boldsymbol{\theta}$ reduces to

$$\begin{aligned} l(\boldsymbol{\theta}) = &- \sum_{i \in F} \log\left[1 + \nu_i^2\, \sinh^2(z_i)\right] + \sum_{i \in F} \log\cosh(z_i) + \sum_{i \in F} (\tau_i - 1) \log\left\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu_i\, \sinh(z_i)]\right\} \\ &+ \sum_{i \in F} \log(\tau_i \nu_i) - \sum_{i \in F} \log(\sigma_i \pi) + \sum_{i \in C} \log\left(1 - \left\{\frac{1}{2} + \frac{1}{\pi}\arctan\left[\nu_i\, \sinh\left(z_i\right)\right]\right\}^{\tau_i}\right). \end{aligned} \tag{17}$$

The MLE $\widehat{\boldsymbol{\theta}}$ of the vector $\boldsymbol{\theta}^T = (\mu, \sigma, \nu, \tau)$ of unknown parameters can be evaluated by maximizing the log-likelihood (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 (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 Stasinopouls (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 (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 $(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ is $N_p(\mathbf{0}, I(\boldsymbol{\theta})^{-1})$, where $I(\boldsymbol{\theta})$ is the expected information matrix. The multivariate normal $N_p(\mathbf{0}, \ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}})^{-1})$ distribution can be used to construct approximate confidence intervals for the individual parameters, where $\ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}})$ is the observed information matrix. Following (16), 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 $T = \hat{\beta}_{12}/\sqrt{\ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}})^{-1}_{\beta_{12}}} \sim t_{(n-p-1)}$, and $\ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}})^{-1}_{\beta_{12}}$ 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 $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$.

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 (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 (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 $\log(T_1), \ldots, \log(T_n)$ are generated from the ESC distribution using the qf (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 $C_1, \ldots, C_n$, 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 (15) for survival censored data and the optim algorithm is used for

maximizing the total log-likelihood function (17).

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 1.

Table 1: The AEs, biases and MSEs based on 2,000 simulations for the location ESC regression model when β_{01}=1, β_{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.

		$n = 50$				$n = 100$		
κ	Parameter	AE	Bias	MSE	Parameter	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 4 by considering the AEs given in Table 1 for $n = 100$, and considering the maximum and minimum values of the generated x_i variable.

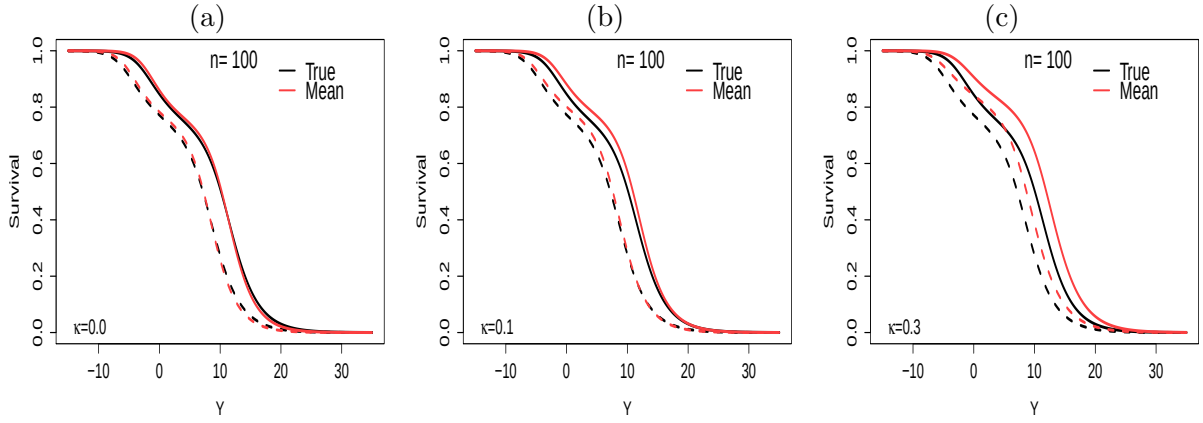


Figure 4: Some ESC survival functions at the true parameter values and at the AEs obtained in Table 1, considering $n = 100$ for the maximum and minimum of x_i when: (a) κ=0 ; (b) κ=0.1; (c) κ=0.3.

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 2.

Table 2: The AEs, biases and MSEs based on 2,000 simulations of the ESC regression model when β_{01}=0.5, β_{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.

		$n = 50$				$n = 100$		
κ	Parameter	AE	Bias	MSE	Parameter	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 5 and the AEs are listed in Table 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 1 and 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 2, should be 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

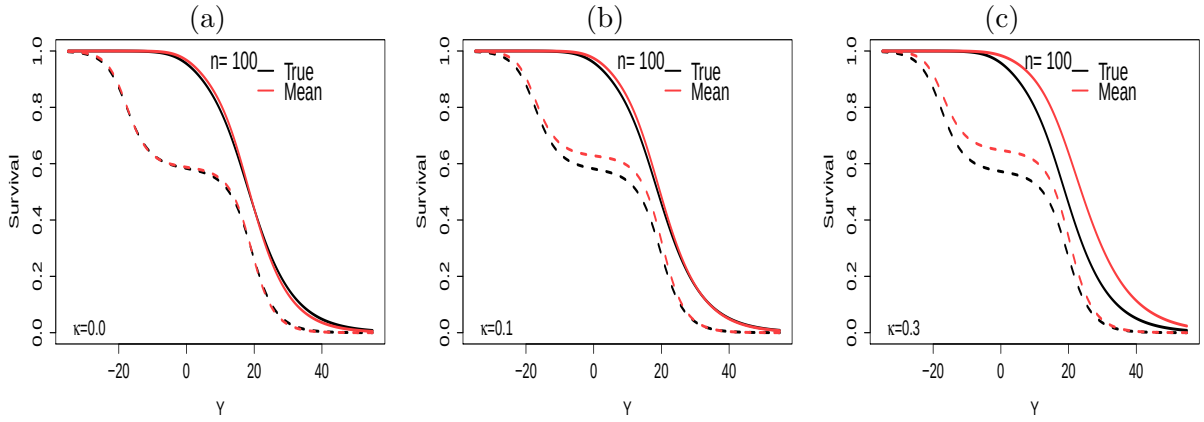


Figure 5: Some ESC survival functions at the true parameter values and at the AEs obtained in Table 2, considering $n = 100$ for the maximum and minimum of x_i when: (a) κ=0 ; (b) κ=0.1; (c) κ=0.3.

regression models, the variances and MSEs increase when the censoring percentage increases. This fact can be noted in Figures 4 and 5.

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

	log-Weibull			normal		
Parameter	$\rho = 0\%$	$\rho = 10\%$	$\rho = 30\%$	$\rho = 0\%$	$\rho = 10\%$	$\rho = 30\%$
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(-)

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.

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 ith case from the data set. The case-deletion model for model (15) is given by

$$y_l = \mu_l + \sigma_l\, z_l, \qquad l = 1, \ldots, n, \quad l \neq i, \tag{18}$$

where the random error Z_l has a density function $f(z_l; \nu_l, \tau_l)$ given in (7). Of course, not always the explanatory variables will be modeling all parameters. For example, if we consider the class of location in (18), the case-deletion model reduces to

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

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 ith case deleted. For model (18), the log-likelihood function of $\boldsymbol{\theta}$ is denoted by $l_{(i)}(\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}_{(i)}^T = (\hat{\boldsymbol{\mu}}_{(i)}^T, \hat{\boldsymbol{\sigma}}_{(i)}^T, \hat{\boldsymbol{\nu}}_{(i)}^T, \hat{\boldsymbol{\tau}}_{(i)}^T)$ be the MLE of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$ from $l_{(i)}(\boldsymbol{\theta})$. To assess the influence of the ith case on the MLE $\hat{\boldsymbol{\theta}}^T = (\hat{\boldsymbol{\mu}}^T, \hat{\boldsymbol{\sigma}}^T, \hat{\boldsymbol{\nu}}^T, \hat{\boldsymbol{\tau}}^T)$, the basic idea is to compare the difference between $\hat{\boldsymbol{\theta}}_{(i)}$ and $\hat{\boldsymbol{\theta}}$. If deletion of a case seriously influences the estimates, for example changing the inference, more attention should be given to that case. Hence, if $\hat{\boldsymbol{\theta}}_{(i)}$ is far from $\hat{\boldsymbol{\theta}}$, then the ith case is regarded as an influential observation. A first measure of the global influence is defined as the standardized norm of $(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})$, known as the generalized Cook distance, defined by

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

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

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

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 (15). If likelihood displacement $LD(\boldsymbol{\omega}) = 2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})\}$ is used, where $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ denotes the MLE under the perturbed model, the normal curvature for $\boldsymbol{\theta}$ in the

direction $\mathbf{d}$, $\|\mathbf{d}\| = 1$, is given by $C_{\mathbf{d}}(\boldsymbol{\theta}) = 2|\mathbf{d}^T\boldsymbol{\Delta}^T\ddot{\mathbf{L}}_{\theta\theta}^{-1}\boldsymbol{\Delta}\mathbf{d}|$, where $\boldsymbol{\Delta}$ is a $p \times n$ matrix that depends on the perturbation scheme, whose elements are given by $\Delta_{vi} = \partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})/\partial\theta_v\partial\omega_i$, $i = 1, \ldots, n$ and $v = 1, \ldots, p$, evaluated at $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega}_0$, and $\boldsymbol{\omega}_0$ is the no perturbation vector. We can also calculate normal curvatures $C_{\mathbf{d}}(\boldsymbol{\mu})$, $C_{\mathbf{d}}(\boldsymbol{\sigma})$, $C_{\mathbf{d}}(\boldsymbol{\nu})$ and $C_{\mathbf{d}}(\boldsymbol{\tau})$ 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 $\mathbf{B} = -\boldsymbol{\Delta}^T\ddot{\mathbf{L}}_{\theta\theta}^{-1}\boldsymbol{\Delta}$ and the index plots of $C_{\mathbf{d}_i}(\boldsymbol{\mu})$, $C_{\mathbf{d}_i}(\boldsymbol{\sigma})$, $C_{\mathbf{d}_i}(\boldsymbol{\nu})$ and $C_{\mathbf{d}_i}(\boldsymbol{\tau})$, 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 ith position. Thus, the curvature in the direction $\mathbf{d}_i$ takes the form $C_i = 2|\boldsymbol{\Delta}_i^T\ddot{\mathbf{L}}_{\theta\theta}^{-1}\boldsymbol{\Delta}_i|$, where $\boldsymbol{\Delta}_i^T$ denotes the ith row of $\boldsymbol{\Delta}$. 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 ith observation is $U_i = \sum_{k=1}^{n_1} \lambda_k e_{ki}^2$, where $\{(\lambda_k, \mathbf{e}_k)|k = 1, \ldots, n\}$ are the eigenvalue-eigenvector pairs of $\mathbf{B}$ with $\lambda_1 \geq \cdots \geq \lambda_{n_1} \geq \lambda_{n_1+1} = \cdots = \lambda_n = 0$ and $\{\mathbf{e}_k = (e_{k1}, \ldots, e_{kn})^T\}$ 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 (15) and log-likelihood function (17), for three perturbation schemes, the matrix

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

6.2.1 Case-weight perturbation

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

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

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

6.2.2 Response perturbation

Since the values of y_i have different variances, they require a scaling of the perturbation vector $\boldsymbol{\omega}$ 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

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

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

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

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

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

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 $r_{M_i} = \delta_i + \log[S(y_i;\hat{\boldsymbol{\beta}})]$, where $\delta_i = 0, 1$ denotes a censored and uncensored observation, respectively, and $S(y_i;\hat{\boldsymbol{\beta}})$ denotes the survival function of Y discussed in Section 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

$$r_{M_i} = \begin{cases} 1 - \hat{\tau}_i\,\log(2\pi) + \log\left[(2\pi)^{\hat{\tau}_i} - \left\{\pi + 2\arctan[\hat{\nu}_i\,\sinh(\hat{z}_i)]\right\}^{\hat{\tau}_i}\right] & \text{if } i \in F \\ -\hat{\tau}_i\,\log(2\pi) + \log\left[(2\pi)^{\hat{\tau}_i} - \left\{\pi + 2\arctan[\hat{\nu}_i\,\sinh(\hat{z}_i)]\right\}^{\hat{\tau}_i}\right] & \text{if } i \in C, \end{cases} \tag{19}$$

where $\hat{z}_i = (y_i - \hat{\mu}_i)/\hat{\sigma}_i$, $\mu_i = \hat{\beta}_{01} + \ldots + x_1[i,p_1+1]\hat{\beta}_{p_1 1}$, $\sigma_i = \exp(\hat{\beta}_{02} + \ldots + x_2[i,p_2+1]\hat{\beta}_{p_2 2})$, $\nu_i = \exp(\hat{\beta}_{03} + \ldots + x_3[i,p_3+1]\hat{\beta}_{p_3 3})$ and $\tau_i = \exp(\hat{\beta}_{04} + \ldots + x_4[i,p_4+1]\hat{\beta}_{p_4 4})$. 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

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

where r_{M_i} is defined in equation (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 $u_i = F(y_i|\hat{\boldsymbol{\theta}}_i)$.

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 $\hat{r}_{ij}$, $j = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, n$;
- 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.

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 subsections 7.1 and 7.2 are available available on the Web at `http://goo.gl/zANZuz` and `http://goo.gl/ZBf8R8`, respectively.

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 ith 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 (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 6.

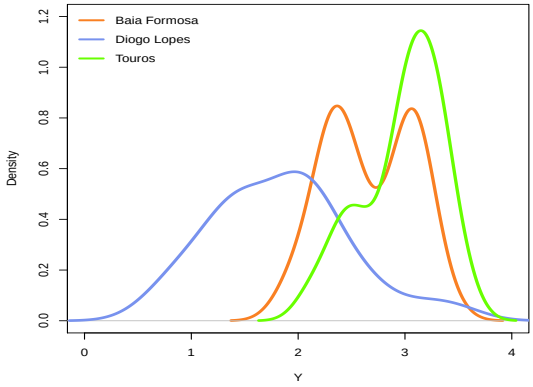


Figure 6: The empirical density of Y in the different regions.

Next, we present results by fitting the model

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

where z_i has density function (7) and the model parameters are defined by

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

Table 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 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.

7.1.1 Global influence analysis

Here, we compute the case deletion measures $GD_i(\boldsymbol{\theta})$ and $LD_i(\boldsymbol{\theta})$ for the shrimp data. The results of such influence measure index plots are displayed in Figure 7. We may note that the 62th observation is a possible influential observation.

Table 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

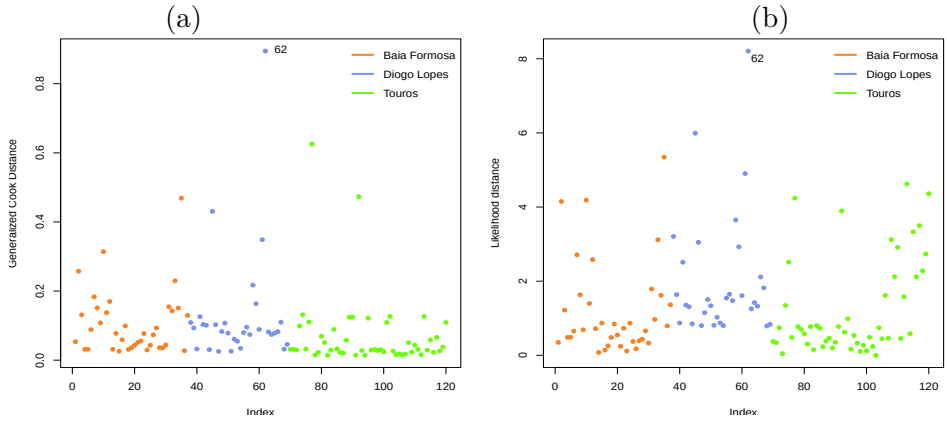


Figure 7: Index plots for $\boldsymbol{\theta}$: (a) $GD_i(\boldsymbol{\theta})$ (Generalized Cook's Distance) and (b) $LD_i(\boldsymbol{\theta})$ (Likelihood Distance).

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 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.

Response perturbation

Next, the influence of perturbations in the observed times is analyzed. Here, we adopt the U_i measure instead of $\mathbf{d}_{max}$ because the first eight eigenvalues are large. Figure 9 displays the index plot of the U_i measure and the total local influence C_i.

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 9 represent the measurements $y_{105} = 2.89$ and $y_{107} = 2.88$ of the Touro region. Combining with the plots of Figure 6, we can note that these two shrimps stabilize the growth of the density.

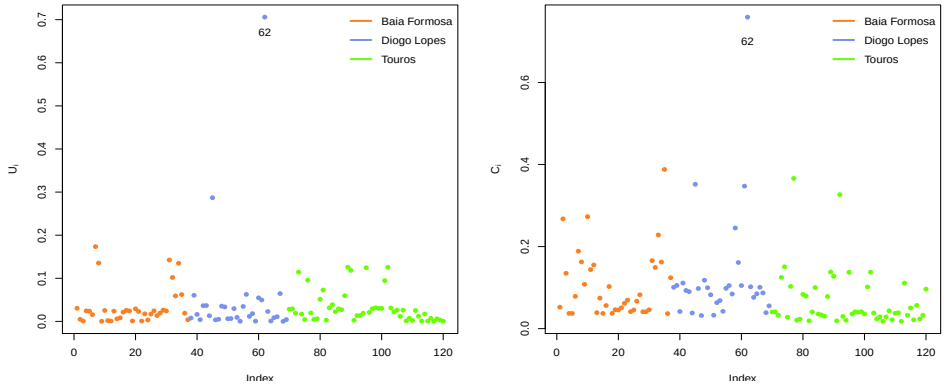


Figure 8: Index plots for $\boldsymbol{\theta}$ (case-weight perturbation): (a) $\mathbf{d}_{max}$ and (b) total local influence.

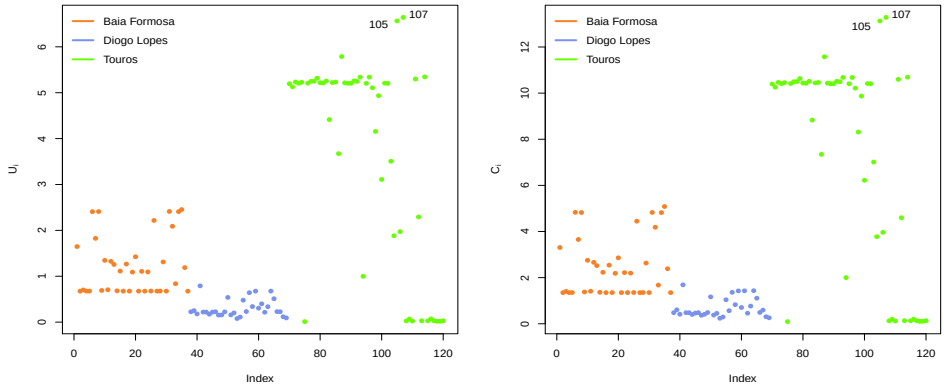


Figure 9: Index plots for $\boldsymbol{\theta}$ (response perturbation): (a) $\mathbf{d}_{max}$ and (b) total local influence.

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 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 11 for different regions. We conclude that the Exp-ESC regression model provides a very good fit to the shrimp data.

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 São 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

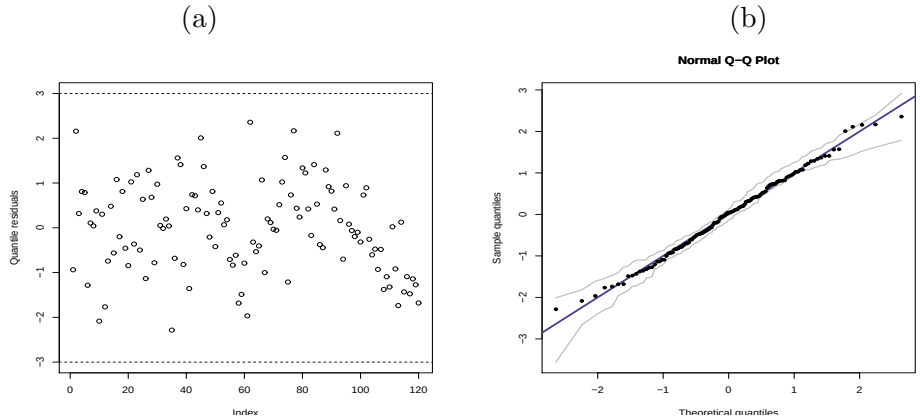


Figure 10: (a) Index plot of the quantile residuals for the shrimp data. (b) Normal probability plot with envelope for the quantile residuals from the fitted ESC regression model to the shrimp data.

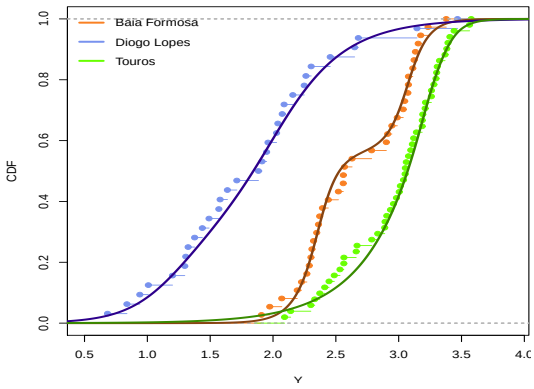


Figure 11: Estimated cumulative fitted values from the ESC fitted model to the shrimp data.

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 $i = 1, \ldots, 172$, has density function $f(z_i; \nu, \tau_i)$ given by (7) and the model parameters are given by

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

Table 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 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

7.2.1 Global influence analysis

Here, we compute the case deletion measures $GD_i(\boldsymbol{\theta})$ and $LD_i(\boldsymbol{\theta})$ for the entomology data. The results of such influence measure index plots are displayed in Figure 12. Based on these plots, we note that the cases 92 and 133 are possibly influential observations.

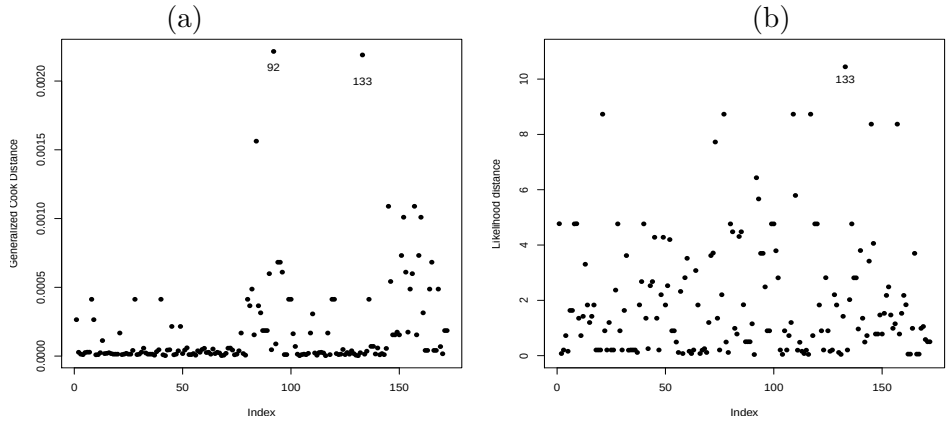


Figure 12: Index plots for $\boldsymbol{\theta}$: (a) $GD_i(\boldsymbol{\theta})$ (Generalized Cook's Distance) and (b) $LD_i(\boldsymbol{\theta})$ (Likelihood Distance).

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 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.

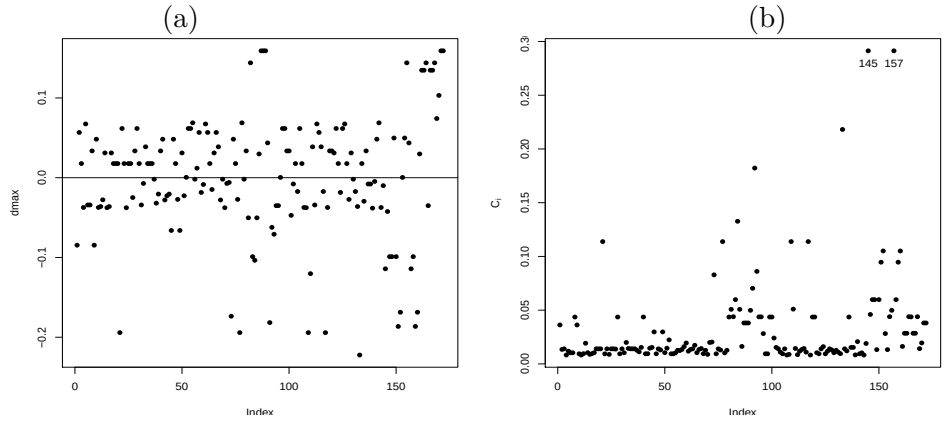


Figure 13: Index plots for $\boldsymbol{\theta}$ (case-weight perturbation): (a) $\mathbf{d}_{max}$ and (b) total local influence.

Response perturbation

The influence of perturbing the observed response Y will be analyzed. The value for the maximum curvature obtained is $C_{\mathbf{d}_{max}} = 10.41$. Figure 14 display the index plots for $\mathbf{d}_{max}$ and total local influence C_i. We may conclude that the observations 96 and 153 are possible influential points.

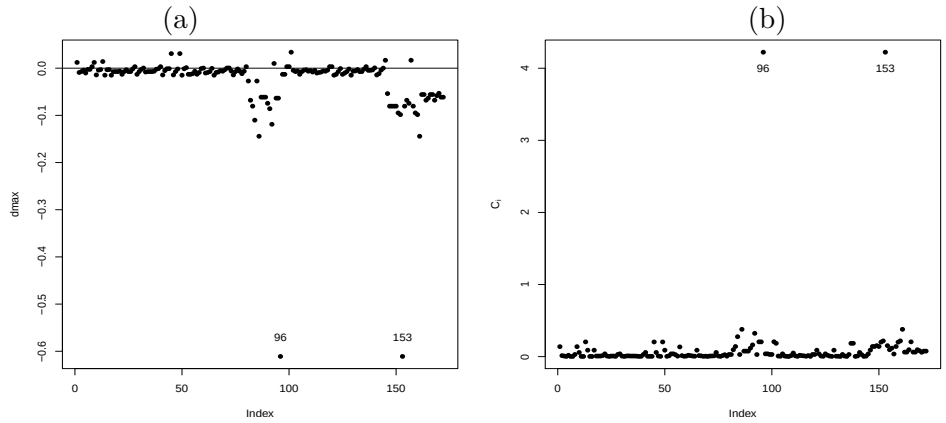


Figure 14: Index plots for $\boldsymbol{\theta}$ (response perturbation): (a) $\mathbf{d}_{max}$ and (b) total local influence.

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.

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 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.

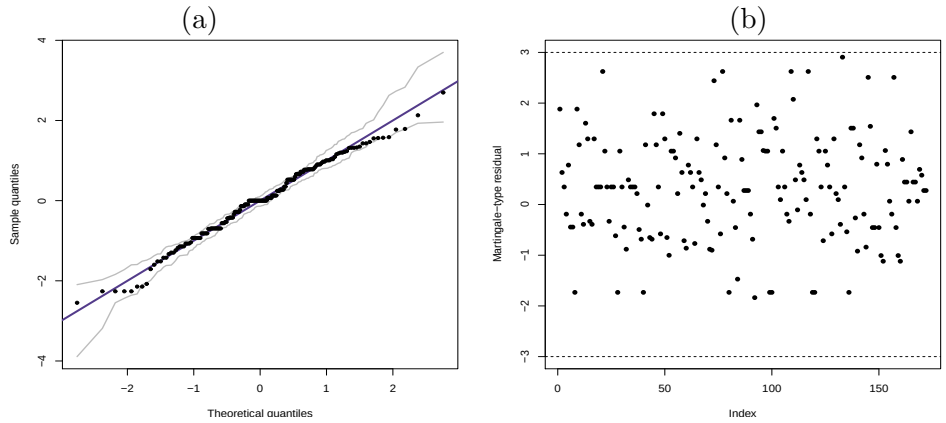


Figure 15: (a) Normal probability plot with envelope for the martingale-type residual r_{D_i} from the fitted ESC regression model to the entomology data. (b) Index plot of the martingale-type residual r_{D_i} for the entomology data.

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 16 for the different groups. We may conclude from the plots that the ESC regression model provides a suitable fit to the entomology data.

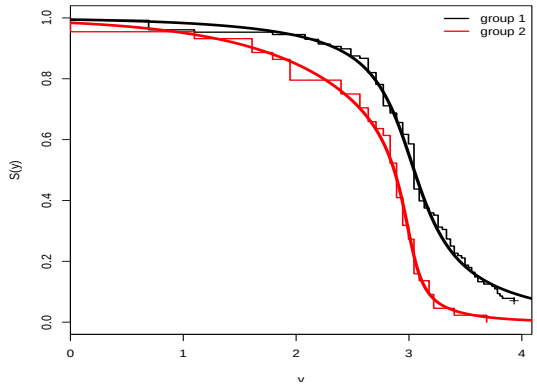


Figure 16: Estimated and empirical survival functions for the entomology data.

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 re-

gression 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 GAMLLS script available in the paper.

Appendix: 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 da Cruz, J.N. (2016). The odd Birnbaum-Saunders regression model with applications to lifetime data. *Journal of Statistical Theory and Practice,* **10**, 780-804.

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 biomètrica 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.