A flexible semiparametric regression model for bimodal, asymmetric and censored data
Peer-reviewed author version


DOI: 10.1080/02664763.2017.1369499
Handle: http://hdl.handle.net/1942/26549
A flexible semiparametric regression model for bimodal, asymmetric and censored data

Thiago G. Ramires  
Department of Mathematic, Federal University of Technology - Paraná, Brazil  
Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-Biostat),  
University of Hasselt, Belgium;

Edwin M.M. Ortega  
Department of Exact Sciences, University of SÃ£o Paulo, Brazil

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

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

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

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 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, which 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: Censored data; diagnostics; P-splines; regression models; semiparametric model.
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 parametric and semiparametric models to extend well-known distributions and to provide flexibility in modeling data has 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 and 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 and Voudouris et al. (2012) showed an application of the Box-Cox power exponential distribution for modeling the location, scale and skewness parameters using P-splines bases.

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. (2015) introduced the exponentiated log-sinh Cauchy (ELSC) distribution to accommodate various shapes of skewness, kurtosis and bi-modality. Based on the log-transformation $Y = \log(T)$, where $T$ has the ELSC distribution, we define the exponentiated sinh Cauchy (ESC) linear regression model in the generalized additive model for location, scale and shape framework (Rigby and Stasinopouls, 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

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

and

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

respectively, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are the location and scale parameters, respectively, $\nu > 0$ is the symmetry parameter, characterizing the bimodality of the distribution, and $\tau > 0$ is the skewness parameter. The ESC density (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 2, we define the ESC 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 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 4, which illustrate the flexibility of the proposed class of regression models. Finally, we offer some conclusions in Section 5.

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) introduced 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 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 Stasinopouls, 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.

2.1 Definition

Let \( \theta^T = (\mu, \sigma, \nu, \tau) \) denote the vector of parameters of the pdf (1). We consider independent observations \( y_i \)'s conditional on \( \theta_i \) (for \( i = 1, 2, \ldots, n \)), having 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, is defined by

\[
y_i = \mu_i + \sigma_i z_i, \quad i = 1, \ldots, n,
\]

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

\[
f(z; \nu, \tau) = \frac{\tau \nu}{\pi} \frac{\cosh(z)}{\nu^2 \sinh^2(z)} + 1 \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan[\nu \sinh(z)] \right\}^{\tau-1}, \quad \text{for} \ z \in \mathbb{R}.
\] (4)

Random errors are statistical fluctuations in the measured data due to the precision limitations of the measurement device. In practice, it is generally considered that \( Z_i \sim N(0, 1) \), but this causes some limitations, since we are assuming that the random errors have a unimodal symmetrical distribution. Plots of the density function (4) for selected parameter values are displayed in Figure 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.

We can define the vector of parameters $\theta$ using appropriate link functions as

$$
\theta = \begin{bmatrix} \mu & \sigma & \nu & \tau \end{bmatrix}^T = \begin{bmatrix} g_1(X_1\beta_1) \\ g_2(X_2\beta_2) \\ g_3(X_3\beta_3) \\ g_4(X_4\beta_4) \end{bmatrix}
$$

or

$$
\theta_i = \begin{bmatrix} \mu_i & \sigma_i & \nu_i & \tau_i \end{bmatrix}^T = \begin{bmatrix} g_1(X_1[i, 2]\beta_{11} + \cdots + X_1[i, p_1 + 1]\beta_{p_1 1}) \\ g_2(X_2[i, 2]\beta_{12} + \cdots + X_2[i, p_2 + 1]\beta_{p_2 2}) \\ g_3(X_3[i, 2]\beta_{13} + \cdots + X_3[i, p_3 + 1]\beta_{p_3 3}) \\ g_4(X_4[i, 2]\beta_{14} + \cdots + X_4[i, p_4 + 1]\beta_{p_4 4}) \end{bmatrix},
$$

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 function, $g_k(\cdot)$, for $k = 2, 3, 4$, are known positive continuously differentiable function containing values of the explanatory variables, $\beta_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{p_k k})^T$ is a parameter vector of length $(p_k + 1)$ and $X_k$ is a known model matrix of order $n \times (p_k + 1)$, whose elements are given by $X_k[i, p_k]$. 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 $p_1 = p_4 = 0$, 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.

### 2.2 Nonparametric additive functions

The ESC GAMLSS model allows to model the distribution parameters $\mu$, $\sigma$, $\nu$ and $\tau$ as linear, nonlinear parametric, nonparametric (smooth) function of the explanatory variables and/or random-effects terms. The parametric regression structure (5) can be extended to semiparametric structure as

$$
\theta = \begin{bmatrix} \mu & \sigma & \nu & \tau \end{bmatrix}^T = \begin{bmatrix} g_1(X_1\beta_1 + \sum_{j=1}^{J_1} h_{1j}(x_{j1})) \\ g_2(X_2\beta_2 + \sum_{j=1}^{J_2} h_{2j}(x_{j2})) \\ g_3(X_3\beta_3 + \sum_{j=1}^{J_3} h_{3j}(x_{j3})) \\ g_4(X_4\beta_4 + \sum_{j=1}^{J_4} h_{4j}(x_{j4})) \end{bmatrix},
$$

for $\theta_i = \begin{bmatrix} \mu_i & \sigma_i & \nu_i & \tau_i \end{bmatrix}^T = \begin{bmatrix} g_1(X_1[i, 2]\beta_{11} + \cdots + X_1[i, p_1 + 1]\beta_{p_1 1}) \\ g_2(X_2[i, 2]\beta_{12} + \cdots + X_2[i, p_2 + 1]\beta_{p_2 2}) \\ g_3(X_3[i, 2]\beta_{13} + \cdots + X_3[i, p_3 + 1]\beta_{p_3 3}) \\ g_4(X_4[i, 2]\beta_{14} + \cdots + X_4[i, p_4 + 1]\beta_{p_4 4}) \end{bmatrix}$. 

Figure 1: Plots of the density function (4) for several values of $\tau$: (a) $\nu = 0.3$; (b) $\nu = 0.8$. 

![Figure 1](image-url)
where \( h_{jk}(x_{jk}) \) are smooth functions of the explanatory variables \( x_{jk} \), for \( k = 1, 2, 3, 4 \) and \( j = 1, \ldots, J_k \), and \( J_k \) is the number of selected explanatory variables to be modeled with the smooth functions. 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 in this paper by B-spline basis functions in the explanatory variables, where the coefficients of the basis functions are penalized to guarantee sufficient smoothness. Rigby and Stasinopouls (2005) proved that each smoothing function \( h_{jk}(\cdot) \) can be expressed as a random effects model, i.e., \( h_{jk}(Z_{jk}) = Z_{jk}\gamma_{jk} \), where \( Z_{jk} \) is an \( n \times q_{jk} \) matrix representing the B-spline basis design matrix and \( \gamma_{jk} \) 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).

2.3 Estimation

In this subsection, 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 script of the \textit{R} software using the computational codes implemented by the first author and available at \url{https://goo.gl/hAIcBF}. The maximization algorithms used are the RS and CG procedures, described by Rigby and Stasinopouls (2005) and Stasinopoulos and Rigby (2007), and available in the documentation of the GAMLSS script.

- **Parametric model**

Consider a sample of \( n \)-independent observations \( y_1, \ldots, y_n \). For the parametric ESC regression model (5), the log-likelihood for the model parameters \( \theta = (\mu^T, \sigma^T, \nu^T, \tau^T)^T \) reduces to

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

- **Survival model**

Consider a sample \( y_1, \ldots, y_n \) of \( n \) independent observations, where each random response is defined by \( y_i = \min \{ \log(t_i), \log(c_i) \} \). Let \( F \) and \( C \) be the sets of individuals for which \( y_i \) is the log-lifetime or log-censoring, respectively. We consider non-informative censoring such that the observed lifetimes and censoring times are independent. The log-likelihood function for the vector of parameters \( \theta = (\mu^T, \sigma^T, \nu^T, \tau^T)^T \) from model (3) has the form

\[
l(\theta) = \sum_{i \in F} l_i(\theta) + \sum_{i \in C} l_i^{(c)}(\theta),
\]

where \( l_i(\theta) = \log[f(y_i)] \),
\( l_i^{(c)}(\theta) = \log[S(y_i)] \), \( f(y_i) \) is the pdf (1) and \( S(y_i) = 1 - F(y_i) \) is the survival function (2), for \( i = 1, \ldots, n \). Therefore, the log-likelihood function for \( \theta \) reduces to

\[
l(\theta) = \sum_{i \in F} \log \left( \frac{\tau_i \mu_i}{\pi \sigma_i} \right) + \sum_{i=1}^{n} \log \left[ \cos \left( \frac{y_i - \mu_i}{\sigma_i} \right) \right] - \sum_{i \in F} \log \left[ 1 + \nu_i^2 \sinh^2 \left( \frac{y_i - \mu_i}{\sigma_i} \right) \right] + \\
(\tau_i - 1) \sum_{i \in F} \log \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[ \nu_i \sinh \left( \frac{y_i - \mu_i}{\sigma_i} \right) \right] \right\} + \\
\sum_{i \in C} \log \left[ 1 - \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[ \nu_i \sinh \left( \frac{y_i - \mu_i}{\sigma_i} \right) \right] \right\}^{\tau_i} \right],
\]

where \( \mu_i, \sigma_i, \nu_i \) and \( \tau_i \) are defined in equation (5).

The log-likelihood (8) can also be maximized in the GAMLSS script using the additional package \texttt{gamlss.cens} to determine numerically the observed information corresponding to the censored observations.

- **Semiparametric model**

Considering the semiparametric model (6), for fixed smoothing parameters \( \lambda_{jk} \), the fixed and random effects \( \beta \) and \( \gamma \), respectively, are estimated by maximizing the penalized log-likelihood function

\[
l_p = \sum_{i \in F} \log \left( \frac{\tau_i \nu_i}{\pi \sigma_i} \right) + \sum_{i=1}^{n} \log \left[ \cos \left( \frac{y_i - \mu_i}{\sigma_i} \right) \right] - \sum_{i \in F} \log \left[ 1 + \nu_i^2 \sinh^2 \left( \frac{y_i - \mu_i}{\sigma_i} \right) \right] + \\
(\tau_i - 1) \sum_{i \in F} \log \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[ \nu_i \sinh \left( \frac{y_i - \mu_i}{\sigma_i} \right) \right] \right\} + \\
\sum_{i \in C} \log \left[ 1 - \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan \left[ \nu_i \sinh \left( \frac{y_i - \mu_i}{\sigma_i} \right) \right] \right\}^{\tau_i} \right] - \frac{1}{2} \sum_{k=1}^{J_k} \sum_{j=1}^{J_k} \lambda_{jk} \gamma_{jk}^T P_{jk} \gamma_{jk},
\]

where \( P_{jk} \) is a symmetric matrix that may depend on a vector of smoothing parameters (see Rigby and Stasinopoulous, 2005). The score functions relative to the likelihood (9) are given by

\[
U^T(\theta) = \frac{\partial l_p}{\partial \theta} = \begin{bmatrix} U_{\beta_1}, U_{\gamma_{j1}}, U_{\beta_2}, U_{\gamma_{j2}}, U_{\beta_3}, U_{\gamma_{j3}}, U_{\beta_4}, U_{\gamma_{j4}} \end{bmatrix},
\]

where elements are given in the Appendix. For each smoothing term selected, and any of the parameters of the ESC distribution, there is one smoothing parameter \( \lambda \) associated with it. The smoothing parameters can be fixed or estimated from the data. We adopt the penalized quasi-likelihood (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 \texttt{pb}(., .). Details about the PQL method can be founded in Rigby and Stasinopoulous (2014) and other methods like Generalized Akaike information criterion (GAIC) and local Generalized Cross validation (GCV) are also available to use with the implemented codes (see Stasinopoulous and Rigby, 2007). One important thing to recognize when fitting a smooth nonparametric term is that the resulting coefficients of the smoothing terms and their standard errors should not be interpreted.
2.4 Model strategy

In this subsection, 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 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 \( GD = -2 l(\hat{\theta}) \), \( l(\hat{\theta}) \) is the total log-likelihood function, \( df \) is the total effective degrees of freedom of the fitted model and \( k \) is a constant. Note that the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are special cases of the GAIC(\( k \)) statistic, corresponding to \( k = 2 \) and \( k = \log(n) \), respectively. The model with the smallest value of these statistics is then selected.

Let \( df_{\mu}, df_{\sigma}, df_{\nu}, \) and \( df_{\tau} \) be the effective degrees of freedom used for modelling \( \mu, \sigma, \nu \) and \( \tau \), 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 \texttt{edfAll()} function in the \texttt{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 \( \mu, \sigma, \nu \) and \( \tau \). Here, we consider the strategy described by Voudouris \textit{et al.} (2012), which is defined in the GAMLSS manual as strategy A (see the \texttt{stepGAICA1.A} function in the GAMLSS script). Let \( \chi \) be the selection of all terms available for consideration, where \( \chi \) could contain both linear and smoothing terms. Then, for all terms in \( \chi \) and for fixed distribution, the strategy is given as follows:

1. use a forward selection procedure to select an appropriate model for \( \mu \) with \( \sigma, \nu \) and \( \tau \) fitted as constants;
2. use a forward selection procedure to select an appropriate model for \( \sigma \) given the model for \( \mu \) obtained in (1) and for \( \nu \) and \( \tau \) fitted as constants;
3. use a forward selection procedure to select an appropriate model for \( \nu \) given the model for \( \mu \) and \( \sigma \) obtained in (2) with \( \nu \) fitted as a constant;
4. use a forward selection procedure to select an appropriate model for \( \tau \) given the model for \( \mu, \sigma \) and \( \tau \) obtained in (3);
5. use a backward selection procedure to select an appropriate model for $\nu$ given the model for $\mu$, $\sigma$ and $\tau$ obtained in (4);
6. use a backward selection procedure to select an appropriate model for $\sigma$, given the model for $\mu$, $\nu$ and $\tau$ obtained in (5);
7. use a backward selection procedure to select an appropriate model for $\mu$ given the model for $\sigma$, $\nu$ and $\tau$ obtained in (6).

At the end of the steps described above, the final model may contain different subsets from $\chi$ for $\mu$, $\sigma$, $\nu$ and $\tau$.

2.5 Simulation

Let a random variable $Y$ have pdf (1). Inverting $F(y) = u$ in (2), we obtain the quantile function (qf) of $Y$ given by

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

Equation (10) can be used for simulating random variables $y_i \sim \text{ESC}(\mu, \sigma, \nu, \tau)$ by fixing $\mu$, $\sigma$, $\nu$ and $\tau$ and considering that $u$ follows a uniform$(0,1)$ distribution. We can simulate the regression models setting the parameters and using the parametric (5) or semiparametric (6) structure.

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 script. 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 were pioneered by Buuren and Fredriks (2001) which are used to identify regions of an explanatory variable which the model does not have good fit quality. 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 (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 variables, we use the smoothed scatterplot diagram available in the `gamlss.util` script of the R software.

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) is given by

\[ y_l = \mu_l + \sigma_l z_l, \quad l = 1, \ldots, n, \quad l \neq i, \]  

where the random error \( Z_l \) has a density function \( f(z_l; \nu_l, \tau_l) \) given in (4).

In the following, a quantity with subscript “\((-i)\)” refers to the original quantity with the \( i \)th case deleted. For model (11), the log-likelihood function for \( \theta \) is denoted by \( l(-i)(\theta) \). Let \( \hat{\theta}^T_{(-i)} = (\hat{\mu}^T_{(-i)}, \hat{\sigma}^T_{(-i)}, \hat{\nu}^T_{(-i)}, \hat{\tau}^T_{(-i)}) \) be the MLEs of \( \mu, \sigma, \nu \) and \( \tau \) 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 \( \hat{\theta}^T_{(-i)} \) and \( \hat{\theta}^T \). If deletion of a case seriously influences the estimates, by changing the inference, more attention should be given to that case. Hence, if \( \hat{\theta}^T_{(-i)} \) is far from \( \hat{\theta}^T \), then the \( i \)th case is regarded as an influential observation. We work with a popular measure of the difference between \( \hat{\theta}^T_{(-i)} \) and \( \hat{\theta}^T \) 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 (7) for parametric models and (9) for semiparametric models. Note that sometimes, depending on the model and data set, the penalized likelihood can have multiple local maxima, so we suggest to use the \( \hat{\theta} \) as initial values to obtain \( \hat{\theta}^T_{(-i)} \).

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), 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 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 \( y_1, \ldots, y_n \), are generated from the ESC distribution using the qf (10) and, for each value of \( n \), all results are obtained from 2,000 Monte Carlo replications. We also consider a simulation study for censored data, considering the percentages of noninformative censoring \( \delta = 30\% \). For each scenario, the log-censoring is also generated randomly using the qf by taking \( n = 50 \) and 150. Here, we present and compare the results fitting the semiparametric ESC and normal models,
for each scenario, where the model parameters are defined by

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

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

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 \( \beta_{21} \) and \( \beta_{31} \) for the ESC and normal distributions. The biases and mean squared errors (MSEs) are evaluated and the results are reported in Table 1.

The figures in Table 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 3 displays the fitted and generated terms for the smooth functions under scenarios 1, 2 and 3 for censoring percentage of \( \delta = 0\% \). 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 and negative skewed errors, respectively. The same conclusion follows, for different censoring percentages, based on the results presented in Table 1. Finally, we can conclude that the estimates of smoothing functions are affected when random errors are not properly estimated by the proposed models.

4 Applications

In this section, we provide two applications to real data to illustrate the flexibility of the semiparametric ESC regression models. The analysis 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.
Table 1: The biases and MSEs of the ESC and normal parametric and semiparametric regression models based on 2,000 simulations for each scenario, censoring percentages and different sample sizes.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>δ</th>
<th>n</th>
<th>Parameter</th>
<th>Bias</th>
<th>MSE</th>
<th>Parameter</th>
<th>Bias</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0%</td>
<td>50</td>
<td>$\beta_{21}$</td>
<td>0.169</td>
<td>18.174</td>
<td>$\beta_{21}$</td>
<td>0.122</td>
<td>24.859</td>
</tr>
<tr>
<td></td>
<td></td>
<td>150</td>
<td>$\beta_{31}$</td>
<td>0.015</td>
<td>30.01</td>
<td>$\beta_{31}$</td>
<td>0.014</td>
<td>1.736</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>$\beta_{21}$</td>
<td>0.014</td>
<td>0.507</td>
<td>$\beta_{21}$</td>
<td>0.029</td>
<td>3.399</td>
</tr>
<tr>
<td>2</td>
<td>0%</td>
<td>50</td>
<td>$\beta_{21}$</td>
<td>0.038</td>
<td>1.671</td>
<td>$\beta_{21}$</td>
<td>0.039</td>
<td>2.023</td>
</tr>
<tr>
<td></td>
<td></td>
<td>150</td>
<td>$\beta_{31}$</td>
<td>0.025</td>
<td>0.340</td>
<td>$\beta_{31}$</td>
<td>0.009</td>
<td>0.639</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>$\beta_{21}$</td>
<td>0.007</td>
<td>0.145</td>
<td>$\beta_{21}$</td>
<td>0.010</td>
<td>0.288</td>
</tr>
<tr>
<td>3</td>
<td>0%</td>
<td>50</td>
<td>$\beta_{21}$</td>
<td>0.014</td>
<td>8.133</td>
<td>$\beta_{21}$</td>
<td>0.055</td>
<td>8.325</td>
</tr>
<tr>
<td></td>
<td></td>
<td>150</td>
<td>$\beta_{31}$</td>
<td>0.065</td>
<td>2.020</td>
<td>$\beta_{31}$</td>
<td>0.055</td>
<td>2.103</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>$\beta_{21}$</td>
<td>0.010</td>
<td>0.770</td>
<td>$\beta_{21}$</td>
<td>0.014</td>
<td>1.154</td>
</tr>
<tr>
<td>1</td>
<td>30%</td>
<td>50</td>
<td>$\beta_{21}$</td>
<td>0.135</td>
<td>5.380</td>
<td>$\beta_{21}$</td>
<td>0.191</td>
<td>4.203</td>
</tr>
<tr>
<td></td>
<td></td>
<td>150</td>
<td>$\beta_{31}$</td>
<td>0.041</td>
<td>5.566</td>
<td>$\beta_{31}$</td>
<td>0.051</td>
<td>10.709</td>
</tr>
<tr>
<td>2</td>
<td>30%</td>
<td>50</td>
<td>$\beta_{21}$</td>
<td>0.125</td>
<td>2.380</td>
<td>$\beta_{21}$</td>
<td>0.091</td>
<td>3.203</td>
</tr>
<tr>
<td></td>
<td></td>
<td>150</td>
<td>$\beta_{31}$</td>
<td>0.025</td>
<td>0.487</td>
<td>$\beta_{31}$</td>
<td>0.019</td>
<td>0.938</td>
</tr>
<tr>
<td>3</td>
<td>30%</td>
<td>50</td>
<td>$\beta_{21}$</td>
<td>0.078</td>
<td>18.035</td>
<td>$\beta_{21}$</td>
<td>0.054</td>
<td>18.279</td>
</tr>
</tbody>
</table>

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 7,482 males and has in total 212 missing values for the explanatory variables, which are then removed. To reduce the computational time of this analysis (approximately 75 hours for the full sample), we consider only the observations of the North corresponding 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
Figure 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 \( n = 300 \) observations for the (a) ESC and (b) normal models.

Before fitting the regression models, as the preliminary analysis, we note in Figure 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.

Next, we present results of the semiparametric ESC and normal models using the steps proposed
Table 2: MLEs of the model parameters for the body mass data, the corresponding SEs (given in parentheses) and the GD, AIC and BIC statistics.

<table>
<thead>
<tr>
<th>Model</th>
<th>Estimates</th>
<th>GD</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESC(µ, σ, ν, τ)</td>
<td>2.738</td>
<td>0.126</td>
<td>0.981</td>
<td>2.862</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.055)</td>
<td>(0.215)</td>
<td>(2.761)</td>
</tr>
<tr>
<td>Normal(µ, σ)</td>
<td>2.894</td>
<td>0.153</td>
<td></td>
<td>-829.7</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.027)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 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.

The model parameters are defined by

$$
\text{ESC} \left\{ \begin{array}{l}
\mu_i = \beta_{01} + pb_{11}(X_{1i}, df), \\
\sigma_i = \exp(\beta_{02} + \beta_{12}x_{1i}), \\
\nu_i = \exp(\beta_{03}) \text{ and } \\
\tau_i = \exp(\beta_{04} + \beta_{14}x_{1i}); \\
\end{array} \right.
$$

$$
\text{Normal} \left\{ \begin{array}{l}
\mu_i = \beta_{01} + pb_{11}(X_{1i}, df) \text{ and } \\
\sigma_i = \exp(\beta_{02} + \beta_{12}x_{1i}). \\
\end{array} \right.
$$

Table 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.

The results presented in Table 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 3, we present in Figure 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 5(b) indicates failure for modelling the kurtosis and skewness for the normal model. We may note in Figures 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, which appear to be consistent with the effects presented in Figure 4(b). Figure 6(a) indicates that the log of the body mass index increases quickly until one year age, and
Table 3: MLEs of the parameters and the approximate SEs from the fitted semiparametric ESC and normal models to the body mass data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semiparametric ESC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{01}$</td>
<td>2.745</td>
<td>0.017</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$\beta_{11}(x_{11},10.35)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{02}$</td>
<td>-3.002</td>
<td>0.089</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.030</td>
<td>0.006</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$\beta_{03}$</td>
<td>-0.234</td>
<td>0.088</td>
<td>0.008</td>
</tr>
<tr>
<td>$\beta_{04}$</td>
<td>-0.049</td>
<td>0.210</td>
<td>0.813</td>
</tr>
<tr>
<td>$\beta_{14}$</td>
<td>0.053</td>
<td>0.021</td>
<td>0.013</td>
</tr>
<tr>
<td>Semiparametric normal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{01}$</td>
<td>2.738</td>
<td>0.005</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$\beta_{11}(X_{11},9.55)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{02}$</td>
<td>-2.433</td>
<td>0.043</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.015</td>
<td>0.003</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$\beta_{03}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\beta_{04}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\beta_{14}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

GD = −1606.0 AIC = −1575.3 BIC = −1501.3

then decreases at a slower rate until 6 years age, and after that increases until 21 years age. The plots in Figure 6(b)-(c) reveal that the variability and skewness of $Y$ increase when $x_1$ increases.

Next, we compute the case deletion $LD_i(\theta)$ measure for the body mass data. Figures 7(a)-(b) reveals the influence measure index plots and the values of $Y$ against $X_1$ with some possible influential points highlighted, respectively. Based on 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 5(c)). In fact, the case 263 has the lowest value of $Y$ and the case 447 has the highest value of $Y$ for the range $10 < X_1 < 15$.

Finally, Figure 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 semiparametric ESC regression model has unimodal shapes with null and positive skewness, e.g. for $x_{1i} = 7$ and $x_{1i} = 19$, respectively. Figure 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.

4.2 Eruption data

In this subsection, we present 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 $x_{1i}$ the waiting time for the eruption. This data set can be obtained using 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 4 gives the MLEs (and the corresponding SEs in parentheses) of the model parameters and the values of the statistics GD, AIC and BIC for the fitted models. Figure 9(a) provides the plots of the histogram of the current data and the fitted densities of the ESC and normal models. Table 4 and Figure 9(a) indicate that the ESC model provides a good fit to these data.
Figure 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: The fitted terms (a) $\mu$, (b) $\sigma$ and (c) $\tau$ for the semiparametric ESC regression model given in Table 3.
Figure 7: For body mass data: (a) Index plots for $|LD_i(\theta)|$ and (b) Observed $Y$ against $X_1$.

Figure 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 4: MLEs of the model parameters for the eruption data, the corresponding SEs (given in parentheses) and the GD, AIC and BIC statistics.

<table>
<thead>
<tr>
<th>Model</th>
<th>Estimates</th>
<th>GD</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{ESC}(\mu, \sigma, \nu, \tau)$</td>
<td>1.044 0.076 0.010 1.545</td>
<td>-88.1</td>
<td>-80.1</td>
<td>-65.7</td>
</tr>
<tr>
<td></td>
<td>(0.008) (0.059) (0.306) (0.361)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{Normal}(\mu, \sigma)$</td>
<td>1.185 0.374</td>
<td>237.0</td>
<td>241.0</td>
<td>248.3</td>
</tr>
<tr>
<td></td>
<td>(0.022) (0.062)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To propose the regression models, as a preliminary analysis, we present in Figure 9(b) the values of $Y$ against $X_1$, for which we note that the response variable $Y$ has a nonlinear relationship with the explanatory variable $X_1$, so that the nonlinear models are required.

Using the steps proposed in Section 2.4 to select additive terms, we present and compare the results
Figure 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.

of the semiparametric ESC and normal models, where the model parameters are defined by

$$
\text{ESC} \left\{ \begin{array}{l}
\mu_i = \beta_{01} + pb_{11}(X_{1i}, df), \\
\sigma_i = \exp[\beta_{02} + pb_{12}(X_{1i}, df)], \\
\nu_i = \exp(\beta_{03} + \beta_{13}x_{1i}) \quad \text{and} \quad \tau_i = \exp(\beta_{04} + \beta_{14}x_{1i}); \\
\end{array} \right.
\text{Normal} \left\{ \begin{array}{l}
\mu_i = \beta_{01} + pb_{11}(X_{1i}, df) \quad \text{and} \quad \sigma_i = \exp[\beta_{02} + pb_{12}(X_{1i}, df)].
\end{array} \right.
$$

Table 5 provides the MLEs, their approximate SEs and $p$-values obtained from the fitted semiparametric ESC and normal regression models. The coefficients of the smoothing terms have been omitted to avoid erroneous interpretations.

Table 5: MLEs of the model parameters and the corresponding SEs from the fitted semiparametric ESC and normal regression models to the eruption data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>SE</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{01}$</td>
<td>1.798</td>
<td>0.017</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$pb_{11}(x_{1i}, 11.21)$</td>
<td>0.967</td>
<td>0.334</td>
<td>0.004</td>
</tr>
<tr>
<td>$pb_{12}(x_{1i}, 5.47)$</td>
<td>3.293</td>
<td>0.517</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$\beta_{03}$</td>
<td>-0.049</td>
<td>0.007</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$\beta_{04}$</td>
<td>5.136</td>
<td>0.340</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$\beta_{05}$</td>
<td>-0.080</td>
<td>0.004</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

$GD = -507 \quad AIC = -465 \quad BIC = -391$

To verify the adequacy and the assumptions of the proposed models in Table 5, we present in Figure 10 the worm plots for four ranges of $X_1$ and the index plots for the quantile residuals. Figure 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 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.

![Graphs showing quantile residuals and index plots](image)

Figure 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.

The partial effects of $X_1$ in the parameters of the semiparametric ESC regression model are presented in Figure 11. Figure 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 additional, Figure 11(b) reveals that the variability of the log of eruption times decays rapidly for $x_1 > 60$. Figures 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 12(a). In Figure 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 10, indicating again the flexibility of the new model.

Finally, Figure 13(a) reveals the semiparametric ESC regression model fitted to the eruption data
Figure 11: The fitted terms for (a) $\mu$, (b) $\sigma$, (c) $\nu$ and (d) $\tau$ for the semiparametric ESC regression model.

Figure 12: For eruption data: (a) Index plots for $|LD_i(\theta)|$ and (b) Observed $Y$ against $X_1$.

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 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 regression could be chosen as the best model to the current data.

5 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 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 script and available from the authors. Two real data sets are used to
Figure 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 \(X_1\).

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.

**Acknowledgements**

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

**Appendix: Score functions**

Let \(U^T(\theta) = \partial l_p/\partial \theta = [U_{\beta_1^1}, U_{\beta_2^1}, U_{\gamma_{rj_1}^1}, U_{\beta_3^1}, U_{\gamma_{rj_2}^1}, U_{\beta_4^1}, U_{\gamma_{rj_3}^1}, U_{\beta_5^1}, U_{\gamma_{rj_4}^1}]\) be the score functions of the likelihood (9), \(\gamma_{rj_1}\) the \(r\)th element of the \(q_{jk}\)-dimensional vector \(\gamma_{jk}\), \(\beta_{l_1 j}\), for \(l_1 = 0, 1, \ldots, p_k\), the \(l\)th element of the vector \(\beta_j\) and \(p_{jk}[r, s]\) the elements of the matrix \(P_{jk}\). The elements of \(\mathbf{U}(\theta)\) are given by

\[
\frac{\partial l(\theta)}{\partial \beta_{l_1 1}} = u_\mu(\beta_{l_1 1}) = \sum_{i \in F} \left[ \frac{\nu_1^2 \sinh(2w_i)}{\sigma_i^2} - \frac{\tanh(w_i)}{\sigma_i} - \frac{\nu_1 \cosh(w_i)}{\pi \sigma B_i K_i} \right] + \sum_{i \in C} \left[ \frac{\nu_1^2}{\sigma_i} \frac{\nu_1 \cosh(w_i) B_i K_i^2}{\pi \sigma_i K_i (1 - B_i^2)} \right],
\]

\[
\frac{\partial l(\theta)}{\partial \gamma_{rj_1}} = u_\mu(\gamma_{rj_1}) - \sum_{j=1}^{J_k} \sum_{s=1}^{q_{jk}} \lambda_{j 1} p_{j 2}[r, s] \gamma_{rj_1},
\]
\[
\frac{\partial l}{\partial \beta_{i2}} = u_\sigma(\beta_{i2}) = \sum_{i \in F} \left[ g_1^{-1}(\sigma_i) \right]_{\beta_{i2}} \left[ -1 - \frac{\nu_i w_i \tanh(w_i)}{\sigma_i} + \frac{\nu_i^2 w_i}{\sigma_i} \sinh(2w_i) - (\tau_i - 1) \frac{\nu_i w_i \cosh(w_i)}{\pi \sigma_i B_i K_i} \right] \\
- \sum_{i \in C} \left[ g_2^{-1}(\sigma_i) \right]_{\beta_{i2}} \frac{\tau_i \nu_i w_i B_i^{r-1} \cosh(w_i)}{\pi \sigma_i K_i (1 - B_i^{r})},
\]

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

\[
\frac{\partial l}{\partial \beta_{i3}} = u_\nu(\beta_{i3}) = \sum_{i \in F} \left[ g_1^{-1}(\nu_i) \right]_{\beta_{i3}} \left[ \frac{1}{\nu_i} - \frac{2\nu_i \sinh^2(w_i)}{K_i} + (\tau_i - 1) \frac{\sinh(w_i)}{\pi B_i K_i} \right] \\
+ \sum_{i \in C} \left[ g_3^{-1}(\nu_i) \right]_{\beta_{i3}} \frac{-\tau_i B_i^{r-1} \sinh(w_i)}{\pi K_i (1 - B_i^{r})},
\]

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

\[
\frac{\partial l}{\partial \beta_{i4}} = u_\tau(\beta_{i4}) = \sum_{i \in F} \left[ g_4^{-1}(\tau_i) \right]_{\beta_{i4}} \left[ \frac{1}{\tau_i} + \log(B_i) \right] + \sum_{i \in C} \left[ g_4^{-1}(\tau_i) \right]_{\beta_{i4}} \frac{-B_i^{r}}{1 - B_i^{r}} \log(B_i)
\]

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

where \[ g_k^{-1}(.), \] for \( k = 1, \ldots, 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 \( w_i = [y_i - \mu_i]/\sigma_i. \)

References


