

Article

# Extended Exponential Regression Model: Diagnostics and Application to Mineral Data

Yolanda M. Gómez <sup>1</sup>, Diego I. Gallardo <sup>1</sup> , Jeremias Leão <sup>2</sup> and Héctor W. Gómez <sup>3,\*</sup> 

<sup>1</sup> Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó 1530000, Chile; yolanda.gomez@uda.cl (Y.M.G.); diego.gallardo@uda.cl (D.I.G.)

<sup>2</sup> Departamento de Estatística, Universidade Federal do Amazonas, Manaus-AM 69001-000, Brazil; jeremias@ufam.edu.br

<sup>3</sup> Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta 1240000, Chile

\* Correspondence: hector.gomez@uantof.cl; Tel.: +56-55-2637-278

Received: 18 October 2020; Accepted: 3 December 2020; Published: 10 December 2020



**Abstract:** In this paper, we reparameterized the extended exponential model based on the mean in order to include covariates and facilitate the interpretation of the coefficients. The model is compared with common models defined in the positive line also reparametrized in the mean. Parameter estimation is approached based on the expectation–maximization algorithm. Furthermore, we discuss residuals and influence diagnostic tools. A simulation study for recovered parameters is presented. Finally, an application illustrating the advantages of the model in a real data set is presented.

**Keywords:** EM algorithm; extended exponential distribution; regression model

## 1. Introduction

Models with positive support have been used a lot in the literature for their usefulness. For example, in areas such as survival analysis, reliability, regression models, among others. In this context, the common models used are the exponential (E), Weibull (W), gamma (G), Birnbaum–Saunders (BS), and generalizations of these distributions, for instance, for limiting cases as illustrated in Fisher and Tippett [1]. A well-known generalization is the one introduced by Gómez et al. [2], named the extended exponential (EE) model with probability density function (PDF)

$$f(y) = \frac{\alpha^2(1 + \beta y) \exp(-\alpha y)}{\alpha + \beta}, \quad y > 0, \alpha > 0 \text{ and } \beta \geq 0. \quad (1)$$

The motivation for this model arises from a mixture between the E distribution with rate  $\alpha$  and the G distribution with shape equal to 2 and rate  $\alpha$ , respectively, where the mixture probability are given by  $\alpha/(\alpha + \beta)$  and  $\beta/(\alpha + \beta)$ , respectively. Another model with a similar motivation is the Lindley (L) distribution Ghitany et al. [3]. The exponentiated generalized EE model has been proposed by Andrade et al. [4]. Rasekhi et al. [5], Rasekhi et al. [6] introduce a generalization and a discrete version of the EE model.

**Remark 1.** The following distribution are particular cases from the  $EE(\alpha, \beta)$  model:

- $EE(\alpha, \beta = 0) = E(\alpha)$ .
- $EE(\alpha, \beta = 1) = L(\alpha)$ .
- $\lim_{\beta \rightarrow +\infty} EE(\alpha, \beta) = G(2, \alpha)$ .

In Remark 1, we can see the flexibility of the model proposed by Gómez et al. [2] since having only two parameters has as special cases three well-known distributions. The mean and variance of the  $EE(\alpha, \beta)$  model are

$$\mathbb{E}(Y) = \frac{\alpha + 2\beta}{\alpha(\alpha + \beta)} \quad \text{and} \quad \text{Var}(Y) = \frac{\alpha^3 + 5\alpha^2\beta + 6\alpha\beta^2 + 2\beta^3}{\alpha^5 + 3\alpha^4\beta + 3\alpha^3\beta^2 + \alpha^2\beta^3},$$

respectively. The rest of the paper proceeds as follows. In Section 2, we introduce a new parameterization of the EE distribution that is indexed by the mean and mixture parameters. Section 3 presents the EE regression model with varying mean and the estimation problem approached via maximum likelihood (ML) estimation via the expectation–maximization (EM) algorithm. In addition, diagnostic measures are discussed. In Section 4, some numerical results of the estimators are presented with a discussion of the obtained results. Furthermore, we discuss an application to real data that shows the usefulness of the proposed model. Concluding remarks are given in Section 5.

## 2. A EE Distribution Parameterized by Its Mean and Mixture Parameters

Regression models are typically obtained to model the mean of a distribution. However, the PDF of the EE distribution given in (1) is indexed by  $\alpha$  and  $\beta$ . In this context, in this section, we considered a new parameterization of the EE distribution in terms of the mean and the mixture proportion of the distribution, say  $\mu > 0$  and  $\pi \in [0, 1]$ , respectively. Consider the parameterization,

$$\begin{aligned} \mu &= \frac{\alpha + 2\beta}{\alpha(\alpha + \beta)} & \alpha &= \frac{2 - \pi}{\mu} \\ \pi &= \frac{\alpha}{\alpha + \beta} & \beta &= \frac{(1 - \pi)(2 - \pi)}{\pi\mu} \end{aligned} \quad \text{or}$$

Under this new parameterization, the PDF in Equation (1), it follows from

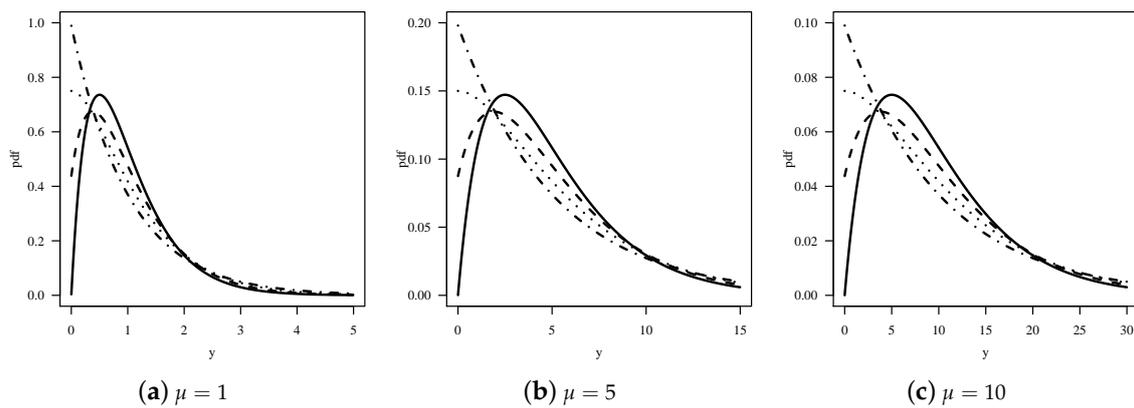
$$f(y; \mu, \pi) = \frac{(2 - \pi)(\mu\pi + (1 - \pi)(2 - \pi)y) \exp[-(2 - \pi)y/\mu]}{\mu^2}, \quad (2)$$

where  $y > 0$ ,  $\mu > 0$  and  $\pi \in [0, 1]$ . Henceforth, we referred to a random variable (RV) with PDF as in (2) as the reparameterized extended exponential model (we denote as  $REE(\mu, \pi)$ ). With this parameterization, based on results in Gómez et al. [2], we have that

$$\mathbb{E}(Y) = \mu \quad \text{and} \quad \text{Var}(Y) = \mu^2[\omega(\pi)]^2,$$

where  $\omega(\pi) = \sqrt{2 - \pi^2}/(2 - \pi)$  is the coefficient of variation (CV). Figure 1 displays some plots of the PDF in (2) for some parameter values. We can notice that the distribution is very flexible and it can be an interesting alternative to other distributions with positive support.

Table 1 gives a summary of the two indices, the skewness and kurtosis for the reparameterized gamma (RGA), reparameterized Birnbaum–Saunders (RBS) and REE distributions, respectively. The interested reader in reparameterized regression models is referred to Santos-Neto et al. [7] and Bourguignon et al. [8,9]. We highlight that in models reparametrized in terms of the mean we can compare the regression coefficients directly.



**Figure 1.** Plots of the REE PDF for indicated  $\mu$  and  $\pi = 0$  (solid line),  $\pi = 0.25$  (dashed line),  $\pi = 0.50$  (dotted line) and  $\pi = 0.90$  (dotdash line).

**Table 1.** Skewness and kurtosis of the RGA, RBS, and REE distributions.

	Skewness	Kurtosis
RGA	$\frac{2}{\sqrt{\phi}}$	$3 + \frac{6}{\phi}$
RBS	$\frac{4(3\phi+11)}{(2\phi+5)^{3/2}}$	$\frac{3(41\phi+186)}{(2\phi+5)^2}$
REE	$\frac{2(2-\pi^3)}{(2-\pi^2)^{3/2}}$	$\frac{3(8-\pi^4-4\pi^2)}{(2-\pi^2)^2}$

### 3. REE Regression Model

Suppose a random sample  $Y_1, \dots, Y_n$  be  $n$  independent RV, where each  $Y_i$ ,  $i = 1, \dots, n$ , follows the PDF given in (2) with mean  $\mu_i$  and mixture proportion parameter  $\pi$ . Suppose the mean parameter of  $Y_i$  satisfies the following functional relation:

$$g_1(\mu_i) = \eta_{1i} = \mathbf{z}_i^\top \boldsymbol{\gamma}, \quad (3)$$

where  $\boldsymbol{\gamma} = (\gamma_0, \dots, \gamma_p)^\top$  is a vector of unknown regression coefficients,  $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$ , with  $p < n$ ,  $\eta_{1i}$  is a linear predictor and  $\mathbf{z}_i = (1, z_{i1}, \dots, z_{ip})^\top$  are observations on  $p$  known regressors, for  $i = 1, \dots, n$ . Furthermore, we assume that the covariate matrices  $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)^\top$  have rank  $p$ . The link functions  $g_1 : \mathbb{R} \rightarrow \mathbb{R}^+$  in (3) must be strictly monotone, positive, and at least twice differentiable, such that  $\mu_i = g_1^{-1}(\mathbf{z}_i^\top \boldsymbol{\gamma})$ , with  $g_1^{-1}(\cdot)$  being the inverse function of  $g_1(\cdot)$ .

Finding the ML estimate of the parameter vector by direct maximization of the log-likelihood can be a hard task. Taking into account the mixture representation of the REE model, we develop an estimation procedure based on the EM algorithm; see Dempster et al. [10] for details about such algorithm.

#### 3.1. EM Algorithm

Considering the mixture representation of the REE distribution, we have

$$\begin{aligned} Y_i | X_i; \mu_i &\sim \text{RGA}(1 + x_i, \mu_i) \\ X_i; \pi &\sim \text{B}(\pi), \end{aligned} \quad (4)$$

where  $\text{RGA}(\mu, \phi)$  with PDF  $f(y; \mu, \phi) \propto y^{\phi-1} \exp\{-\phi y/\mu\}$  and  $\text{B}(\pi)$  denotes the Bernoulli distribution with success probability equal to  $\pi$ . Under this setting,  $\mathbf{D}_{obs} = (\mathbf{Y}, \mathbf{Z})$  represents the observed data,  $\mathbf{X} = (x_1, \dots, x_n)$  represents the unobserved (latent) data and  $\mathbf{D}_{comp} = (\mathbf{Y}, \mathbf{Z}, \mathbf{X})$  denotes the complete data, where  $\mathbf{Y} = (y_1, \dots, y_n)$ ,  $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$  and  $\mathbf{X} = (x_1, \dots, x_n)$ . The complete log-likelihood for  $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \pi)$  is given by

$$\ell_c(\boldsymbol{\psi}; \mathbf{D}_{comp}) \propto \sum_{i=1}^n \left[ (2 - x_i) \log \left( \frac{2 - \pi}{\mu_i} \right) - \frac{(2 - \pi)y_i}{\mu_i} + x_i \log \pi + (1 - x_i) \log(1 - \pi) \right]. \quad (5)$$

Let  $\boldsymbol{\psi}^{(k)}$  be the estimate of  $\boldsymbol{\psi}$  at the  $k$ -th iteration of the EM algorithm and denote the conditional expectation of  $\ell_c(\boldsymbol{\psi}; \mathbf{D}_{comp})$  given the observed data as  $Q(\boldsymbol{\psi} | \boldsymbol{\psi}^{(k)})$ . Therefore,

$$Q(\boldsymbol{\psi} | \boldsymbol{\psi}^{(k)}) \propto - \sum_{i=1}^n \left[ (2 - \tilde{x}_i^{(k)}) \log \mu_i + \frac{(1 + \tilde{x}_i^{(k)})y_i}{\mu_i} - \tilde{x}_i^{(k)} \log \pi - (1 - \tilde{x}_i^{(k)}) \log(1 - \pi) \right], \quad (6)$$

where  $\tilde{x}_i^{(k)} = \mathbb{E}(X_i | \mathbf{D}_{obs}; \boldsymbol{\psi}^{(k)})$ ,  $i = 1, \dots, n$ . The distribution of  $X_i | \mathbf{D}_{obs}; \boldsymbol{\psi}$  is derived in the following proposition.

**Proposition 1.** For the REE model, the distribution of  $X_i | \mathbf{D}_{obs}; \boldsymbol{\psi}$  in the hierarchical representation in (4) is

$$X_i | \mathbf{D}_{obs}; \boldsymbol{\psi} \sim B \left( \frac{\pi \mu_i / (y_i(2 - \pi))}{1 - \pi + \pi \mu_i / (y_i(2 - \pi))} \right), \quad i = 1, \dots, n.$$

**Proof.** The marginal distribution for  $Y_i$  is

$$\begin{aligned} f(y_i | \mathbf{D}_{obs}; \boldsymbol{\psi}) &= \sum_{x_i \in \{0,1\}} f(y_i | X_i = x_i; \mathbf{D}_{obs}; \boldsymbol{\psi}) P(X_i = x_i; \pi) \\ &= 1 - \pi + \pi \mu_i / (y_i(2 - \pi)). \end{aligned}$$

The proof is complete applying the Bayes's theorem for  $P(X_i = x_i | \mathbf{D}_{obs}; \boldsymbol{\psi})$ .  $\square$

**Corollary 1.** The following expected values are directly from Proposition 1.

- $\mathbb{E}(X_i | \mathbf{D}_{obs}; \boldsymbol{\psi}) = \mathbb{E}(X_i^2 | \mathbf{D}_{obs}; \boldsymbol{\psi}) = \frac{\pi \mu_i / (y_i(2 - \pi))}{1 - \pi + \pi \mu_i / (y_i(2 - \pi))}$ ,  $i = 1, \dots, n$ .

In general, the three steps of the Algorithm 1 are:

---

#### Algorithm 1 EM algorithm for REE regression model

---

**E-step.** For  $i = 1, \dots, n$ , compute

$$\tilde{x}_i^{(k)} = \frac{\pi^{(k-1)} \mu_i^{(k-1)} / (y_i(2 - \pi^{(k-1)}))}{1 - \pi^{(k-1)} + \pi^{(k-1)} \mu_i^{(k-1)} / (y_i(2 - \pi^{(k-1)}))}.$$

**M-step I.** Using  $\tilde{\boldsymbol{x}}^{(k)} = (\tilde{x}_1^{(k)}, \dots, \tilde{x}_n^{(k)})$  obtained previously in the E-step, update  $\gamma^{(k)}$  solving the following nonlinear equation for  $\gamma$

$$\sum_{i=1}^n \mathbf{z}_i^\top \left( 1 + \tilde{\boldsymbol{y}}_i^{(k)} \right) = \sum_{i=1}^n \mathbf{z}_i^\top \frac{(1 + \tilde{x}_i^{(k)}) y_i}{\mu_i}, \text{ and;}$$

**M-step II.** Update  $\pi^{(k)}$  as follows

$$\pi^{(k)} = \frac{1}{n} \sum_{i=1}^n \tilde{x}_i^{(k)}.$$


---

The E, M-I and M-II steps are alternated repeatedly until a suitable convergence rule is satisfied, e.g., the difference in successive values of the estimates is less than a tolerance value.

**Remark 2.** When no covariates are available, the M-step I for  $\mu$  can be updated in a closer form as follows:

$$\mu^{(k)} = \frac{\sum_{i=1}^n (1 + \tilde{x}_i^{(k)}) y_i}{\sum_{i=1}^n (1 + \tilde{x}_i^{(k)})}.$$

**Remark 3.** The usual choice for  $g_1(\cdot)$  in (3) is  $g_1(u) = \exp(u)$ .

We carry out inference for  $\psi$  of the REE regression model using asymptotic distribution of the ML estimator  $\hat{\psi}$  obtained by the EM algorithm. This estimator is consistent and has a multivariate normal joint asymptotic distribution with an asymptotic mean  $\psi$  and asymptotic covariance matrix  $\Sigma_\psi$ , respectively, which can be obtained from the corresponding expected information Fisher information matrix  $\mathcal{I}(\psi)$ . Hence, we have

$$\sqrt{n} [\hat{\psi} - \psi] \xrightarrow{D} N_{p+2}(0, \Sigma_{\hat{\psi}} = \mathcal{J}^{-1}(\psi)), \quad (7)$$

as  $n \rightarrow \infty$ , where  $\xrightarrow{D}$  means “convergence in distribution”, and  $\mathcal{J}(\psi) = \lim_{n \rightarrow \infty} \frac{1}{n} \mathcal{I}(\psi)$ . Note that  $\hat{\mathcal{I}}(\hat{\psi})$  is a consistent estimator of the asymptotic covariance matrix of  $\psi$ . According to these results, an approximate  $100 \times [1 - \tau]\%$  confidence region for  $\psi$  obtained from (7) is

$$\mathcal{R} = \{\psi \in \mathbb{R}^{p+2} : [\hat{\psi} - \psi]^\top \hat{\Sigma}_{\hat{\psi}}^{-1} [\hat{\psi} - \psi] \leq \chi_{p+2, 1-\tau}^2\}, \quad 0 < \tau < 1, \quad (8)$$

where  $\chi_{p+2, 1-\tau}^2$  denotes the  $[1 - \tau] \times 100$ th percentile of the chi-squared distribution with  $p + 2$  degrees of freedom and  $\hat{\Sigma}_{\hat{\psi}}$  is an estimate of  $\Sigma_\psi$ .

### 3.2. Diagnostic Analysis

Diagnostic analysis is an important way to detect influential cases and evaluate their effect on the model assumptions. In this subsection, we use the local influence approach to detect influential observations that under small perturbation of the model exert a great influence on the ML estimators. There are basically two approaches to detect influential observations that seriously influence the results of statistical analysis: (A1) the first approach is the case-deletion approach, in which the impact of deleting an observation on the estimators is directly assessed by metrics such as the likelihood distance and Cook’s distance, see Cook [11]; (A2) the second one is to estimate outputs with respect to the model inputs via various minor model perturbations, such as the local influence; see Cook [12]. Zhu and Lee [13] introduced a unified approach for local influence analysis of general statistical models with missing data on the basis of the Q-function for the EM-type algorithm. Alternative methodologies to perform diagnostic analysis can be seen in Bolboaca and Jantschi [14], Jantschi [15].

#### Case Deletion Measures

Case-deletion diagnosis is an approach to detect the effect of dropping the  $i$ th case from the data set. Here, we develop diagnostic measures with the whole vector  $(y_i, x_i)$  deleted and denote these by the subscript “(i)”. The classic measures are the Cook distance and the likelihood displacement. Following this approach, Lee and Xu [16] introduce the analogue to the generalized Cook’s distance ( $GCD_i^c$ ) and likelihood displacement ( $LD_i^c$ ) for the Q-function, which are given by

$$\begin{aligned} GCD_i^c &= (\hat{\psi}_{(i)} - \hat{\psi})^\top [-\ddot{Q}(\hat{\psi} | \psi)] (\hat{\psi}_{(i)} - \hat{\psi}); \\ LD_i^c &= 2\{\ell_c(\hat{\psi}) - \ell_c(\hat{\psi}_{(i)})\}, \end{aligned}$$

where  $\hat{\psi}_{(i)}$  is the maximizer of the Q-function  $Q_{(i)}(\psi | \hat{\psi})$ ,  $i = 1, \dots, n$ .

The Hessian Matrix  $\ddot{Q}_\psi(\psi)$

To obtain the diagnostic measures of the REE regression model, based on the approach of Dempster et al. [10], it is necessary to compute  $\ddot{Q}_\psi(\hat{\psi}) = \partial^2 Q(\psi | \hat{\psi}) / \partial \psi \partial \psi^\top$ , where  $\psi = (\pi, \gamma^\top)^\top$ . It follows from (6) that  $\ddot{Q}_\psi(\hat{\psi})$  have elements given by

$$\frac{\partial^2 Q(\psi | \hat{\psi})}{\partial \gamma \partial \gamma^\top} = - \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^\top \frac{(1 + \tilde{x}_i) y_i}{\mu_i}, \text{ and}$$

$$\frac{\partial^2 Q(\psi | \hat{\psi})}{\partial \pi \partial \pi} = - \sum_{i=1}^n \left[ \frac{\tilde{x}_i^{(k)}}{\pi^2} + \frac{(1 - \tilde{x}_i^{(k)})}{(1 - \pi)^2} \right].$$

### 3.3. Perturbation Schemes

In this subsection, we consider three different perturbation schemes for the REE regression model.

#### 3.3.1. Case Weights Perturbation

Let  $\omega = (\omega_1, \dots, \omega_n)^\top$  an  $n \times 1$  dimensional vector with  $\omega_0 = (1, \dots, 1)^\top$ . Then, the expected value of the perturbed complete-data log-likelihood function (perturbed  $Q$ -function) can be written as  $Q(\psi, \omega | \hat{\psi}) = \sum_{i=1}^n \omega_i \mathbb{E}[\ell_i(\psi | \mathbf{y}_c)] = \sum_{i=1}^n \omega_i Q(\psi | \hat{\psi})$ , where  $Q(\psi | \hat{\psi})$  is given in (6). In this case, the matrix

$$\Delta_{\omega_0} = \left. \frac{\partial^2 Q(\psi, \omega | \hat{\psi})}{\partial \psi \partial \omega^\top} \right|_{\omega = \omega_0} = (\Delta_\gamma^\top, \Delta_\pi^\top)^\top,$$

has elements given by

$$\Delta_\gamma = - \sum_{i=1}^n \omega_0 \mathbf{z}_i^\top \left[ (2 - \tilde{x}_i) \frac{(1 + \tilde{x}_i) y_i}{\mu_i} \right],$$

$$\Delta_\pi = \sum_{i=1}^n \omega_0 \left[ \frac{\tilde{x}_i^{(k)}}{\pi} - \frac{(1 - \tilde{x}_i^{(k)})}{(1 - \pi)} \right].$$

#### 3.3.2. Response Perturbation

We here assume an additive perturbation for the response variables  $\mathbf{y} = (y_1, \dots, y_n)^\top$ , such that  $\mathbf{y}(\omega) = \mathbf{y} + \omega s_z$ , where  $s_z$  is a scale factor equal to the estimated standard deviation of  $\mathbf{y}$ . Then, perturbed  $Q$ -function is given by

$$Q(\psi, \omega | \hat{\psi}) \propto - \sum_{i=1}^n \left[ (2 - \tilde{x}_i^{(k)}) \log \mu_i + \frac{(1 + \tilde{x}_i^{(k)}) y_i(\omega_i)}{\mu_i} - \tilde{x}_i^{(k)} \log \pi - (1 - \tilde{x}_i^{(k)}) \log(1 - \pi) \right].$$

#### 3.3.3. Covariate Perturbation

Here, we are interested in perturbing a specific explanatory variable. Under this condition, we consider an additive perturbation of the explanatory variables by setting  $\mathbf{z}_r(\omega) = \mathbf{z}_r + s_r \omega$ , with  $r = 1, \dots, p$  and  $s_r$  is a scale value typically given by the standard deviation of  $\mathbf{z}_r$ . In this perturbation scheme, the perturbed  $Q$ -function is given by

$$Q(\psi, \omega | \hat{\psi}) \propto - \sum_{i=1}^n \left[ (2 - \tilde{x}_i^{(k)}(\omega_i)) \log \mu_i(\omega_i) + \frac{(1 + \tilde{x}_i^{(k)}(\omega_i)) y_i}{\mu_i(\omega_i)} \right. \\ \left. - \tilde{x}_i^{(k)}(\omega_i) \log \pi - (1 - \tilde{x}_i^{(k)}(\omega_i)) \log(1 - \pi) \right],$$

where  $\mu_i(\omega_i) = \mathbf{x}_r^\top(\omega_i) \gamma$ .

### 3.4. Residual Analysis

In order to check the goodness-of-fit of the REE regression model, we propose two types of residuals for our model, which are the randomized quantile (RQ) and the generalized Cox–Snell (GCS) residuals proposed by [17,18], and given respectively by

$$r_i^{RQ} = \Phi^{-1}(\widehat{S}_Y(y_i; \widehat{\mu}_i, \widehat{\pi})),$$

$$r_i^{GCS} = -\log(\widehat{S}_Y(y_i; \widehat{\mu}_i, \widehat{\pi})), \quad i = 1, \dots, n,$$
(9)

where  $\widehat{\mu}_i = \mathbf{x}^T \widehat{\gamma}$ ,  $\Phi^{-1}$  is the inverse function of the  $N(0, 1)$  cumulative distribution function (CDF) and  $\widehat{S}_Y(y_i; \widehat{\mu}_i, \widehat{\pi})$  is the estimated CDF of the RV given in (2). If the REE model is correctly specified, then the RQ residual has a  $N(0, 1)$  distribution, regardless of the model specification, whereas the GCS residual has an exponential distribution with a parameter equal to one.

### 4. Simulation Study

In this section, we carry out a simulation study to evaluate the performance of the ML estimators of the REE regression model parameters. Here, we consider for each individual two covariates ( $p = 2$ ): one dichotomous, drawn from the  $B(0.5)$  distribution, and one continuous, drawn from the standard normal model. Those covariates were included although  $\mu_i$  following Equation (3). In addition, we consider four values for  $\pi$ : 0.2, 0.5, 0.75, and 0.95, and three combinations for the vector of parameters  $\gamma = (\gamma_0, \gamma_1, \gamma_2)$  related to the covariates:  $(1, 0.5, 0.01)$ ,  $(1, -1, -0.01)$  and  $(2, -0.5, 0.02)$ . We also consider four sample sizes: 50, 100, 200, and 500. Each combination for  $\pi$ ,  $\gamma$ , and  $n$  were replicated 10,000 times. For each scenario and for each parameter, we computed the mean bias, the mean of the estimated standard errors (SE), the root of the mean squared error (RMSE), and the 95% coverage probability (CP) based on the asymptotic distribution for maximum likelihood estimators. Results are presented in Tables 2 and 3. Note that, in all scenarios, the bias for each parameter is acceptable and decreases when the sample size is increasing. The SE and RMSE terms also are closer, reducing their difference when  $n$  increases, suggesting that the variance of estimators is well estimated. Finally, the CP's are closer to the nominal value (95%), especially when  $n$  is increased, which suggests that the distribution of the estimators is well approximated by the normal, even with moderate sample sizes. Simulation and application codes were written in the R programming language, R [19], where they were compiled using the Windows 10 operating system, 16 GB RAM, Intel Core i7 processor, 64 bits.

**Table 2.** Recovery parameters for the REE regression model under different scenarios (cases  $n = 50$  and  $n = 100$ ).

True Values				$n = 50$					$n = 100$			
$\pi$	$\gamma_0$	$\gamma_1$	$\gamma_2$	Estimator	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
0.2	1	0.5	0.01	$\pi$	−0.018	0.200	0.165	0.992	−0.012	0.125	0.124	0.946
				$\gamma_0$	−0.018	0.158	0.161	0.943	−0.010	0.111	0.112	0.950
				$\gamma_1$	−0.007	0.224	0.226	0.949	0.001	0.156	0.155	0.951
				$\gamma_2$	−0.002	0.116	0.119	0.946	−0.001	0.079	0.080	0.949
	1	−1	−0.01	$\pi$	−0.018	0.197	0.167	0.992	−0.011	0.125	0.122	0.951
				$\gamma_0$	−0.017	0.158	0.164	0.941	−0.009	0.111	0.111	0.951
				$\gamma_1$	0.001	0.224	0.228	0.945	0.001	0.157	0.155	0.952
				$\gamma_2$	−0.002	0.115	0.117	0.949	−0.001	0.080	0.081	0.944
	2	−0.5	0.02	$\pi$	−0.021	0.197	0.164	0.989	−0.010	0.125	0.123	0.949
				$\gamma_0$	−0.019	0.158	0.162	0.940	−0.009	0.110	0.113	0.945
				$\gamma_1$	0.003	0.224	0.228	0.945	0.000	0.157	0.158	0.944
				$\gamma_2$	−0.001	0.115	0.117	0.948	0.000	0.079	0.079	0.948

Table 2. Cont.

True Values				$n = 50$					$n = 100$			
$\pi$	$\gamma_0$	$\gamma_1$	$\gamma_2$	Estimator	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
0.5	1	0.5	0.01	$\pi$	-0.060	0.265	0.245	0.881	-0.025	0.198	0.190	0.908
				$\gamma_0$	-0.026	0.175	0.186	0.933	-0.010	0.124	0.127	0.943
				$\gamma_1$	0.004	0.250	0.258	0.938	-0.002	0.176	0.178	0.945
				$\gamma_2$	0.001	0.129	0.135	0.936	0.000	0.090	0.091	0.943
	1	-1	-0.01	$\pi$	-0.062	0.267	0.247	0.882	-0.024	0.198	0.189	0.912
				$\gamma_0$	-0.025	0.175	0.184	0.934	-0.013	0.124	0.126	0.945
				$\gamma_1$	0.002	0.249	0.260	0.932	0.003	0.176	0.178	0.946
				$\gamma_2$	0.000	0.129	0.136	0.937	0.002	0.090	0.091	0.947
	2	-0.5	0.02	$\pi$	-0.062	0.267	0.248	0.883	-0.027	0.198	0.189	0.913
				$\gamma_0$	-0.024	0.175	0.186	0.932	-0.011	0.124	0.127	0.943
				$\gamma_1$	0.002	0.249	0.258	0.938	0.001	0.176	0.179	0.945
				$\gamma_2$	0.000	0.129	0.135	0.938	-0.001	0.090	0.093	0.941
0.75	1	0.5	0.01	$\pi$	-0.136	0.330	0.287	0.857	-0.075	0.267	0.218	0.893
				$\gamma_0$	-0.031	0.187	0.203	0.928	-0.017	0.133	0.140	0.939
				$\gamma_1$	0.002	0.266	0.281	0.934	0.002	0.189	0.199	0.938
				$\gamma_2$	0.001	0.138	0.148	0.935	-0.001	0.096	0.098	0.946
	1	-1	-0.01	$\pi$	-0.138	0.331	0.288	0.853	-0.078	0.266	0.220	0.886
				$\gamma_0$	-0.028	0.186	0.202	0.927	-0.013	0.133	0.138	0.941
				$\gamma_1$	-0.002	0.265	0.283	0.928	-0.002	0.189	0.195	0.939
				$\gamma_2$	0.001	0.138	0.147	0.936	0.000	0.096	0.099	0.944
	2	-0.5	0.02	$\pi$	-0.140	0.328	0.290	0.849	-0.071	0.269	0.216	0.892
				$\gamma_0$	-0.031	0.186	0.203	0.928	-0.015	0.133	0.140	0.935
				$\gamma_1$	0.007	0.265	0.284	0.931	0.006	0.189	0.198	0.938
				$\gamma_2$	0.001	0.138	0.146	0.935	-0.001	0.096	0.100	0.943

Table 3. Recovery parameters for the REE regression model under different scenarios (cases  $n = 200$  and  $n = 500$ ).

True Values				$n = 200$					$n = 500$			
$\pi$	$\gamma_0$	$\gamma_1$	$\gamma_2$	Estimator	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
0.2	1	0.5	0.01	$\pi$	-0.005	0.087	0.088	0.928	-0.001	0.056	0.056	0.943
				$\gamma_0$	-0.004	0.078	0.079	0.949	-0.001	0.049	0.049	0.949
				$\gamma_1$	0.000	0.110	0.111	0.949	-0.001	0.069	0.070	0.951
				$\gamma_2$	0.001	0.056	0.055	0.953	-0.001	0.035	0.035	0.949
	1	-1	-0.01	$\pi$	-0.005	0.087	0.088	0.929	-0.002	0.056	0.056	0.941
				$\gamma_0$	-0.004	0.078	0.078	0.951	-0.002	0.049	0.049	0.947
				$\gamma_1$	0.000	0.110	0.110	0.950	0.000	0.069	0.069	0.950
				$\gamma_2$	0.001	0.055	0.055	0.946	0.001	0.035	0.035	0.947
	2	-0.5	0.02	$\pi$	-0.007	0.087	0.088	0.928	-0.002	0.056	0.055	0.943
				$\gamma_0$	-0.005	0.078	0.079	0.947	-0.002	0.049	0.049	0.951
				$\gamma_1$	0.000	0.110	0.111	0.950	-0.001	0.069	0.069	0.949
				$\gamma_2$	0.000	0.056	0.056	0.949	0.000	0.035	0.035	0.949
0.5	1	0.5	0.01	$\pi$	-0.012	0.141	0.140	0.934	-0.004	0.087	0.086	0.953
				$\gamma_0$	-0.006	0.088	0.089	0.946	-0.003	0.056	0.057	0.946
				$\gamma_1$	0.000	0.124	0.125	0.951	0.001	0.079	0.079	0.950
				$\gamma_2$	0.001	0.063	0.064	0.944	0.001	0.039	0.040	0.945
	1	-1	-0.01	$\pi$	-0.009	0.142	0.139	0.933	-0.004	0.087	0.088	0.950
				$\gamma_0$	-0.005	0.088	0.088	0.948	-0.003	0.056	0.056	0.948
				$\gamma_1$	-0.001	0.125	0.125	0.949	0.001	0.079	0.079	0.948
				$\gamma_2$	0.000	0.063	0.062	0.947	0.000	0.040	0.040	0.947
	2	-0.5	0.02	$\pi$	-0.012	0.140	0.137	0.934	-0.004	0.087	0.087	0.950
				$\gamma_0$	-0.006	0.088	0.089	0.946	-0.003	0.056	0.056	0.947
				$\gamma_1$	-0.001	0.124	0.126	0.946	0.001	0.079	0.078	0.950
				$\gamma_2$	-0.001	0.063	0.063	0.948	0.000	0.040	0.040	0.950

Table 3. Cont.

True Values				$n = 200$					$n = 500$			
$\pi$	$\gamma_0$	$\gamma_1$	$\gamma_2$	Estimator	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
0.75	1	0.5	0.01	$\pi$	-0.041	0.208	0.169	0.913	-0.016	0.141	0.123	0.927
				$\gamma_0$	-0.007	0.095	0.097	0.941	-0.003	0.060	0.062	0.944
				$\gamma_1$	0.001	0.134	0.138	0.941	0.001	0.085	0.086	0.945
				$\gamma_2$	0.000	0.068	0.069	0.945	-0.001	0.043	0.043	0.946
	1	-1	-0.01	$\pi$	-0.039	0.207	0.169	0.914	-0.013	0.142	0.121	0.928
				$\gamma_0$	-0.008	0.095	0.097	0.942	-0.003	0.060	0.061	0.951
				$\gamma_1$	0.002	0.134	0.138	0.940	0.001	0.085	0.085	0.953
				$\gamma_2$	-0.001	0.068	0.069	0.945	0.000	0.043	0.044	0.947
	2	-0.5	0.02	$\pi$	-0.038	0.209	0.168	0.916	-0.014	0.144	0.121	0.931
				$\gamma_0$	-0.007	0.095	0.097	0.943	-0.003	0.060	0.061	0.948
				$\gamma_1$	0.000	0.134	0.137	0.945	-0.001	0.085	0.087	0.946
				$\gamma_2$	0.001	0.068	0.070	0.941	0.000	0.043	0.044	0.945

### 5. Applications

In this section, we present a real data application where the good performance of the REE model is presented over other common models in the literature.

#### 5.1. Exploratory Data Analysis to the Mineral Data Set

We illustrate the proposed methodology by applying it to a real-world mineral data set. These data were obtained from the Department of Mining of the University of Atacama, Chile, to study the concentration of some ores in the soil. This data set corresponds to a sample of 86 measurements of the concentration of the Vanadium and Thorium ores, respectively. We consider a regression model to explain the quantity of vanadium (V) in terms of the quantity of thorium (Th). The study of the concentration of ores in the soil is a matter of public health since it is possible to detect, for example if the tributary water may be contaminated with heavy metals, among others. Similar works can be found in Gómez et al. [20], Bolfarine et al. [21], Olmos et al. [22] and Reyes et al. [23] that verified the concentration of nickel and zinc in the soil.

Table 4 provides a descriptive summary of the observed vanadium concentration that includes median (MD), mean ( $\bar{y}$ ), SD, CV, skewness (CS) and kurtosis (CK), and minimum ( $y_{(1)}$ ) and maximum ( $y_{(n)}$ ) values. The unit of measurement of the concentration (response variable) is parts-per million (ppm). From this table, note the positively skewed nature of the data distribution. The skewed nature is confirmed by the histogram of Figure 2 (left). In addition, Figure 2 (right) indicated some relationship between quantity of vanadium in terms of the quantity of thorium, which motivates the use of the REE regression model to study these variables.

Table 4. Descriptive statistics for mineral data.

$y_{(1)}$	MD	$\bar{y}$	SD	CV	CS	CK	$y_{(n)}$	$n$
1.00	114.50	133.79	104.46	78.82	0.61	2.58	459.00	86

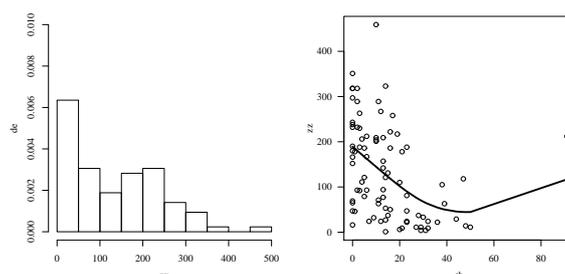


Figure 2. Histogram (left) and Scatterplot with smoothing curve (right) for mineral data.

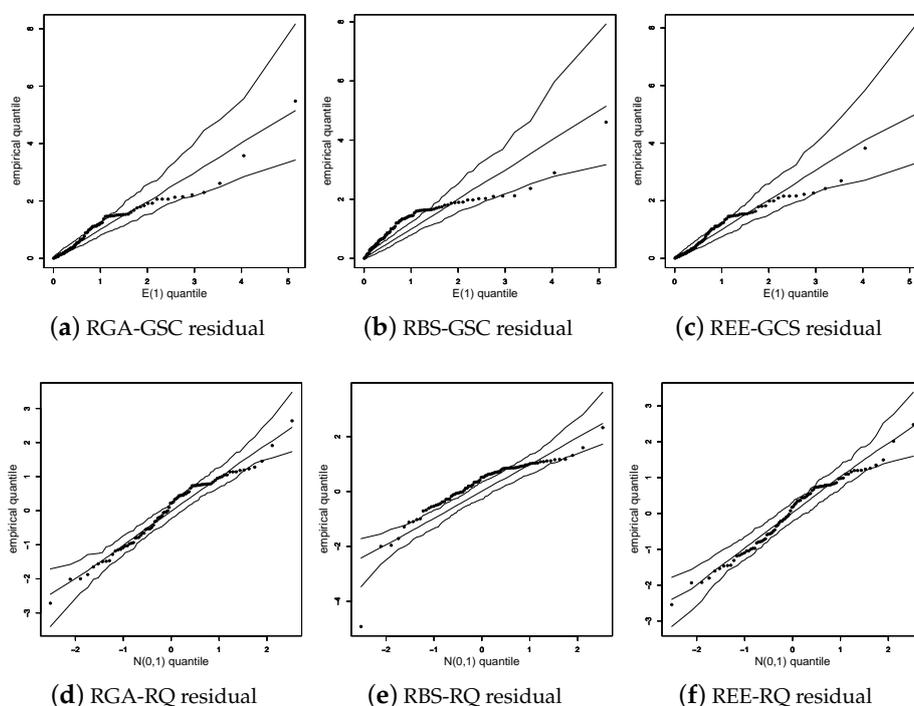
### 5.2. Estimation and Model Checking

We consider the REE regression model with the structure:  $\eta_i = \gamma_0 + \gamma_1 z_{i1}$ . Table 5 provides the estimation and hypothesis testing results for the REE regression model analyzing mineral data. Results of the RGA and RBS regression models are also detailed in this table, as well as their AIC (Akaike [24]), BIC (Schwarz [25]), and log-likelihood values.

**Table 5.** ML estimates (with estimated asymptotic SEs in parentheses) for the RGA, RBS, and REE regression model for the fit mineral data set.

Parameter	Fitted Models		
	RGA	RBS	REE
$\hat{\gamma}_0$	5.0734 (0.1252)	5.0271 (0.1767)	5.0440 (0.1122)
$\hat{\gamma}_1$	−0.0145 (0.0060)	−0.0172 (0.0063)	−0.0125 (0.0044)
$p$ -value	[0.0148]	[0.0068]	[0.0049]
$\hat{\phi}$	1.1955 (0.1630)	-	-
$\hat{\xi}$	-	0.8888 (0.1376)	-
$\hat{\pi}$	-	-	0.4650 (0.1713)
log-likelihood	−503.3155	−520.3540	−502.5457
AIC	1012.6311	1046.7080	1011.0914
BIC	1019.9941	1054.0711	1018.4544

Figure 3 shows the QQ plots with simulated envelope for the GCS and QS residuals. These plots allow us to check graphically whether the GCS and QS residuals follow the EXP(1) and N(0, 1) distributions or not, respectively. From Figure 3, note that these residuals present a good agreement with their corresponding target distributions.



**Figure 3.** QQ plot with a simulated envelope under the indicated residual and model for mineral data.

### 5.3. Diagnostic Analysis

Based on estimation and model validation results presented previously, we conducted a diagnostic analysis for the REE regression model, the suggested fitted model according to the analysis. Next,

we carry out our diagnostic analysis based on global and local influence. First, Figure 4 presents the case-deletion measures  $GCD_i^c(\psi)$  and  $LD_i^c(\psi)$  discussed in Section 3.2. From this figure,  $LD_i^c(\psi)$  statistic indicates that the cases #25, #44, and #69 are potentially influential. On the other hand, the  $GCD_i^c(\psi)$  statistic does suggest case #44 as potentially influential.

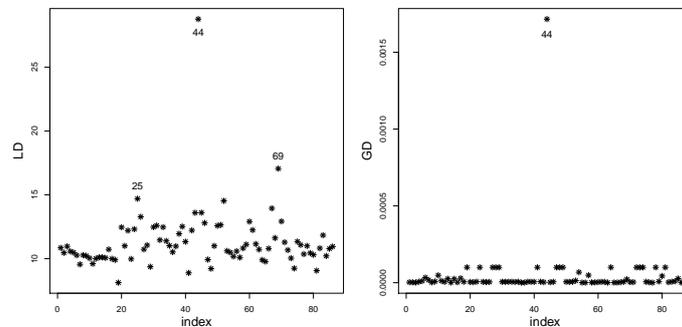


Figure 4. Likelihood displacement (left) and Generalized Cook (right) distance for mineral data.

Index plots of  $C_i$  for  $\pi$  and  $\gamma$  under the case-weight, response, and covariate perturbation schemes are displayed in Figure 5. Note that case #44 is detected as potentially influential on  $\hat{\pi}$  and  $\hat{\gamma}$  under the three perturbation schemes.

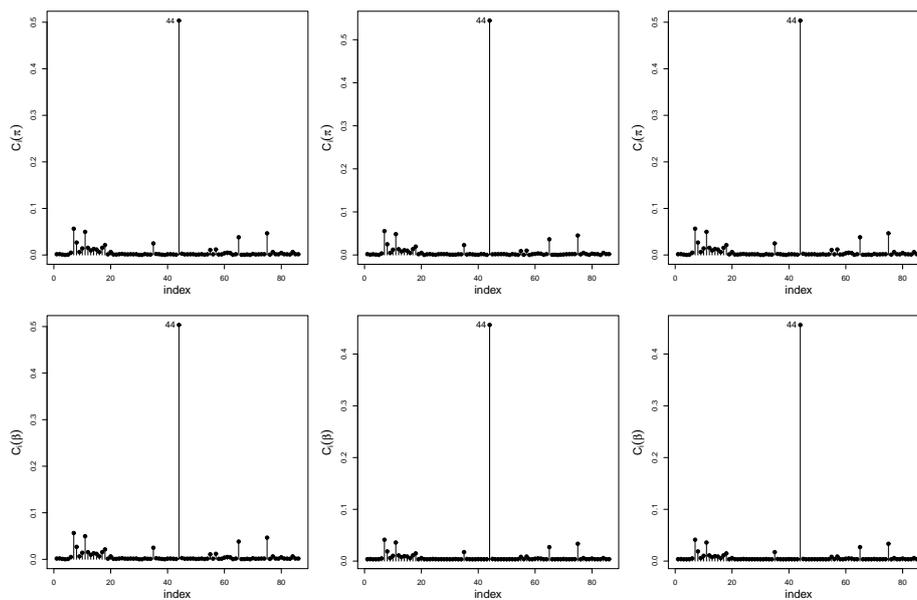


Figure 5. Index plots of  $C_i$  for  $\delta$  under the case-weight (left), response (center) and covariate (right) perturbation schemes with mineral data.

In order to check the impact on the model inference of the detected influential cases, we remove influential cases and reestimating the parameters as well as their corresponding SEs. Table 6 shows the parameter estimates and their corresponding estimated SEs without observation #44. In addition,  $p$ -values are shown for the regression coefficients based on  $t$ -tests.

**Table 6.** ML estimates (with estimated asymptotic SEs in parentheses) for the RGA, RBS, and REE regression model for the fit mineral data set (without observation #44).

Parameter	Fitted Models		
	RGA	RBS	REE
$\hat{\gamma}_0$	5.3248 (0.1240)	5.3260 (0.2020)	5.3008 (0.1276)
$\hat{\gamma}_1$	−0.0399 (0.0066)	−0.0453 (0.0101)	−0.0381 (0.0069)
<i>p</i> -value	[<0.0001]	[<0.0001]	[<0.0001]
$\hat{\phi}$	1.3865 (0.1925)	-	-
$\hat{\xi}$	-	1.0595 (0.1640)	-
$\hat{\pi}$	-	-	0.3056 (0.1548)
log-likelihood	−488.7114	−508.1733	−488.1657
AIC	983.4229	1022.3466	982.3314
BIC	990.7508	1029.6746	989.6594

## 6. Conclusions

In the present paper, the reparameterization of the EE model based on the mean motivated us to propose a regression model for positive data. The maximum likelihood method is employed with the EM algorithm for estimating the model parameters. Application to real data with covariates was presented showed by the AIC and BIC criteria besides the deviance residuals in which the REE model fit better for this data set than the other reparameterizations considered.

The main contribution of this paper is to develop EM algorithms for maximum likelihood estimation as well as to apply Zhu and Lee's approach for case-deletion measures and the local influence diagnostics in the linear regression models with REE errors. Closed-form expressions are obtained for the M and E steps of EM algorithm, for the observed information matrix, for the Hessian matrix  $\ddot{Q}$ , and for the  $\Delta$ matrix under three perturbation schemes. The models can be fitted using standard available software packages, like R (code available upon request).

**Author Contributions:** Conceptualization, Y.M.G., D.I.G. and J.L.; formal analysis, Y.M.G., D.I.G., J.L. and H.W.G.; investigation, Y.M.G., D.I.G., J.L. and H.W.G.; methodology, Y.M.G., D.I.G. and J.L.; software, Y.M.G., D.I.G. and J.L.; supervision, D.I.G. and H.W.G.; validation, Y.M.G., D.I.G. and J.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** The research of Yolanda M. Gómez was support by proyecto DIUDA programa de inserción N° 22367 of the Universidad de Atacama, Chile. The research of H.W. Gómez was supported by Grant PUENTE UA, Chile.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Fisher, R.A.; Tippett, L.H.C. Limiting forms of the frequency distribution of the largest and smallest member of a sample. *Proc. Camb. Philos. Soc.* **1928**, *24*, 180–190. [[CrossRef](#)]
2. Gómez, Y.M.; Bolfarine, H.; Gómez, H.W. A New Extension of the Exponential Distribution. *Colomb. J. Stat.* **2014**, *37*, 25–34. [[CrossRef](#)]
3. Ghitany, M.E.; Atieh, B.; Nadarajah, S. Lindley distribution and its application. *Math. Comput. Simul.* **2008**, *78*, 493–506. [[CrossRef](#)]
4. Andrade, T.A.N.; Bourguignon, M.; Cordeiro, G.M. The exponentiated generalized extended exponential distribution. *J. Data Sci.* **2016**, *14*, 393–414.
5. Rasekhi, M.; Alizadeh, M.; Altun, E.; Hamedani, G.G.; Afify, A.Z.; Ahmad, M. The Modified Exponential Distribution with Applications. *Pak. J. Stat.* **2017**, *33*, 383–398.
6. Rasekhi, M.; Chatrabgoun, O.; Daneshkhah, A. Discrete Weighted Exponential Distribution: Properties and Applications. *Filomat* **2018**, *32*, 3043–3056. [[CrossRef](#)]
7. Santos-Neto, M.; Cysneiros, F.J.; Leiva, V.; Barros, M. Reparameterized Birnbaum–Saunders regression models with varying precision. *Electron. J. Stat.* **2016**, *10*, 2825–2855. [[CrossRef](#)]

8. Bourguignon, M.; Santos-Neto, M.; de Castro, M. A new regression model for positive data. *arXiv* **2018**, arXiv:1804.07734.
9. Bourguignon, M.; Leão, J.; Gallardo, D.I. Parametric modal regression with varying precision. *Biom. J.* **2020**, *62*, 2002–2020. [[CrossRef](#)]
10. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38.
11. Cook, R.D. Detection of influential observation in linear regression. *Technometrics* **1977**, *19*, 15–18.
12. Cook, R.D. Assessment of local influence. *J. R. Stat. Soc. Ser. B* **1986**, *48*, 133–155. [[CrossRef](#)]
13. Zhu, H.-T.; Lee, S.-Y. Local influence for incomplete data models. *J. R. Stat. Soc. Ser. B* **2001**, *63*, 111–126. [[CrossRef](#)]
14. Bolboaca, S.D.; Jantschi, L. The Effect of Leverage and/or Influential on Structure-Activity Relationships. *Comb. Chem. High Throughput Screen.* **2013**, *16*, 288–297. [[CrossRef](#)]
15. Jantschi, L. A Test Detecting the Outliers for Continuous Distributions Based on the Cumulative Distribution Function of the Data Being Tested. *Symmetry* **2019**, *11*, 835. [[CrossRef](#)]
16. Lee, S.-Y.; Xu, L. Influence analyses of nonlinear mixed-effects models. *Comput. Statist. Data Anal.* **2004**, *45*, 321–341. [[CrossRef](#)]
17. Dunn, P.K.; Smyth, G.K. Randomized quantile residuals. *J. Comput. Graph. Stat.* **1996**, *5*, 236–244.
18. Cox, D.R.; Snell, E.J. A general definition of residuals. *J. R. Stat. Soc. Ser. B* **1968**, *30*, 248–265. [[CrossRef](#)]
19. R Development Core Team. *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
20. Gómez, H.W.; Venegas, O.; Bolfarine, H. Skew-symmetric distributions generated by the distribution function of the normal distribution. *Envirometrics* **2006**, *18*, 395–407. [[CrossRef](#)]
21. Bolfarine, H.; Gómez, H.W.; Rivas, L.I. The log-bimodal-skew-normal model. A geochemical application. *J. Chemom.* **2011**, *25*, 329–332. [[CrossRef](#)]
22. Olmos, N.M.; Varela, H.; Gómez, H.W.; Bolfarine, H. An extension of the half-normal distribution. *Stat. Pap.* **2012**, *53*, 875–886. [[CrossRef](#)]
23. Reyes, J.; Barranco-Chamorro, I.; Gallardo, D.I.; Gómez, H.W. Generalized Modified Slash Birnbaum–Saunders Distribution. *Symmetry* **2018**, *10*, 724. [[CrossRef](#)]
24. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [[CrossRef](#)]
25. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [[CrossRef](#)]

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).