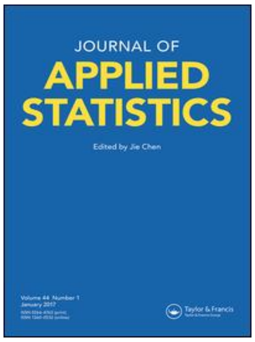


Taylor & Francis
Taylor & Francis Group

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

# A new regression model for bimodal data and applications in agriculture

Julio Cezar Souza Vasconcelos, Gauss Moutinho Cordeiro, Edwin Moises Marcos Ortega & Édila Maria de Rezende

**To cite this article:** Julio Cezar Souza Vasconcelos, Gauss Moutinho Cordeiro, Edwin Moises Marcos Ortega & Édila Maria de Rezende (2020): A new regression model for bimodal data and applications in agriculture, Journal of Applied Statistics, DOI: 10.1080/02664763.2020.1723503

**To link to this article:** https://doi.org/10.1080/02664763.2020.1723503

Published online: 05 Feb 2020.

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2020.1723503

# A new regression model for bimodal data and applications in agriculture

Julio Cezar Souza Vasconcelos[a], Gauss Moutinho Cordeiro[b], Edwin Moises Marcos Ortega[a] and Édila Maria de Rezende[c]

[a]ESALQ, Universidade de São Paulo, Piracicaba, Brazil; [b]UFPE, Universidade Federal de Pernambuco, Recife, Brazil; [c]UFLA, Universidade Federal de Lavras, Lavras, Brazil

**ABSTRACT**

We define the odd log-logistic exponential Gaussian regression with two systematic components, which extends the heteroscedastic Gaussian regression and it is suitable for bimodal data quite common in the agriculture area. We estimate the parameters by the method of maximum likelihood. Some simulations indicate that the maximum-likelihood estimators are accurate. The model assumptions are checked through case deletion and quantile residuals. The usefulness of the new regression model is illustrated by means of three real data sets in different areas of agriculture, where the data present bimodality.

**ARTICLE HISTORY**
Received 8 July 2019
Accepted 25 January 2020

**KEYWORDS**
Agriculture data; bimodal data; exponential Gaussian distribution; regression model; simulation study

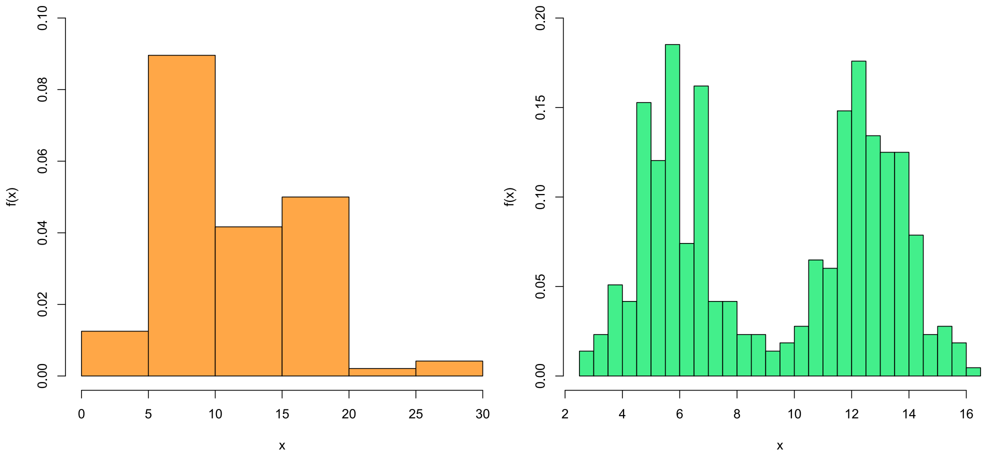
## 1. Introduction

The normal (Gaussian) distribution is used to model many phenomena in almost all areas. It is adequate for real data when most of the data are near to the mean. On the other hand, the exponential is a continuous distribution with positive support. It is one of the simplest probabilistic models used to describe time to failure.

If $Y_1 \sim N(\mu, \sigma^2)$ and $Y_2 \sim \text{Exp}(\nu)$, where $\nu = E(Y_2)$, and $Y_1$ and $Y_2$ are independent random variables, then the sum $Y = Y_1 + Y_2$ has the *exponential Gaussian* (ExGa) distribution, say $Y \sim \text{ExGa}(\mu, \sigma^2, \nu)$. Some results were reported for the ExGa distribution. For example, [27] implemented this distribution in **R** software (GAMLSS), [7] proved that it may provide better fits for some classes of phenomena including intermitotic time and protein expression variability data. Further, [11] used the ExGa distribution for reconstruction of chromatographic peaks, [18] applied it in experiments to measure response item and [29] used this distribution for integrated extended time-lapse automated imaging to quantify the dynamics of cell proliferation. All these papers consider unimodal data, but in some situations, this assumption does not hold. For example, we consider the following datasets:

- The data are related to the index of germination speed of tomato seeds. The research was developed at the Central Seed Laboratory of the Federal University of Lavras, Lavras, MG, Brazil (see Figure 1(a)).

**CONTACT** Julio Cezar Souza Vasconcelos juliocezarvasconcelos@hotmail.com



(a) (b)

**Figure 1.** Histograms.

- Another data set consists of the degrees Brix (a measure of the density or sugar concentration of solutions) of *yacon* (a tuber native to Peru) (see Figure 1(b)).

Figure 1(a,b) displays the existence of a bimodal data distribution. These data sets are analyzed in this paper in the application section. Our first objective is to define a new distribution, called the *odd log-logistic exponential Gaussian* (OLLExGa), to model data with two modes (bimodal). In many practical situations, the response variable is affected by several explanatory variables, such as temperature, radiation, sulfurgran, ascorbic acid, germination index, among others. The regression model that provides a better fit tends to produce more precise estimates for the quantities of interest.

Recently, some studies of regressions have been published in different contexts. For example, [21] introduced the heteroscedastic odd log-logistic generalized gamma regression for censored data, [10] studied a zero-spiked regression models generated by gamma random variables with application in the resin oil production and [22] considered a generalized odd log-logistic flexible Weibull regression with applications in repairable systems, [26] proposed the odd log-logistic generalized inverse Gaussian with real estate data regression, among others. Further, [9] defined the G family of continuous distributions with mathematical properties, characterizations and regression modeling, [13] presented the odd power Lindley generator of probability distributions with properties, characterizations and regression modeling, [14] introduced the Weibull Marshall–Olkin family with regression and application to censored data and [12] proposed a new flexible lifetime model with log-location regression, properties and applications.

Based on these surveys, our second objective is to construct a regression based on the OLLExGa distribution to model bimodal data by considering a classic analysis and with different applications in agriculture. The inferential part is carried out using asymptotic maximum-likelihood estimators (MLEs). Some Monte Carlo simulation studies are performed to verify the accuracy of the OLLExGa regression by means of the variance and mean squared error. We check the model assumptions and detect possible influential or extreme observations that can cause distortions in the results of the fitted regression. An efficient way to detect these observations, called case deletion or global influence, was proposed by [2]. We introduce quantile residuals (qrs) to check the regression assumptions and carry out simulation studies to evaluate their empirical distribution when the data are bimodal. We draw envelope plots as a measure of the goodness-of-fit. Our research can be summarized in the following contributions:

- First, we present the OLLExGa distribution to model bimodal data.
- Second, based on the OLLExGa distribution, we propose a regression with two systematic components to model bimodal data. There are no classic models for bimodal data in the literature.
- Third, we present diagnostic and residual analysis to verify all assumptions of the new regression.
- Finally, we present three applications where the main motivation is the presence of bimodality in these data. We emphasize that in the first application, the researcher responsible for the execution of the experiment provides all final interpretations of these analyses. She even emphasized that the normal regression cannot be adopted for these data. In these terms, we are sure that our proposed regression can be used not only in the area of agriculture, but regression may be used in other areas. We focus on agricultural applications.

In Section 2, we define the OLLExGa distribution and display some plots. In Section 3, we propose the OLLExGa regression and investigate the accuracy of the MLEs from several simulations. In Section 4, we define qrs for the fitted regression and some diagnostic measures. We also provide a simulation study to check the normal approximation for these residuals. Three applications to real data in agriculture area in Section 5 confirm the flexibility of the OLLExGa distribution and its associated regression model. Section 6 ends with some conclusions.

## 2. The model definition

It is important to have extended forms of classic distributions in many applied areas such as agriculture data modeling. We adopt the parametrization of the ExGa distribution used in the GAMLSS library [27] in **R**. The cumulative distribution function (cdf) and probability density function (pdf) of the ExGa distribution are

$$G_{\mu,\sigma,\nu}(y) = \frac{1}{\nu}\int_0^y \exp\left(\frac{\mu - t}{\nu} + \frac{\sigma^2}{2\,\nu^2}\right)\Phi\left(\frac{t-\mu}{\sigma} - \frac{\sigma}{\nu}\right)\mathrm{d}t, \quad y \in \mathbb{R}, \tag{1}$$

and

$$g_{\mu,\sigma,\nu}(y) = \frac{1}{\nu}\exp\left(\frac{\mu - y}{\nu} + \frac{\sigma^2}{2\nu^2}\right)\Phi\left(\frac{y-\mu}{\sigma} - \frac{\sigma}{\nu}\right), \tag{2}$$

respectively, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are the mean and standard deviation of the normal distribution, $\nu > 0$ is the mean of the exponential variable and $\Phi(\cdot)$ is the standard normal cumulative function.

Let $W \sim \mathrm{ExGa}(\mu, \sigma, \nu)$ be a random variable having density function (2). The moment generating function (mgf) of $W$ is $M_W(t) = (1 - \nu t)^{-1} \exp(\mu t + \sigma^2 t^2/2)$. It can be checked from $M_W(t)$ that the ExGa distribution converges to the normal distribution when $\nu$ goes to zero. By differentiating $M_W(t)$, the mean, variance, skewness and kurtosis of $W$ are

$$E(W) = \mu + \nu, \quad V(W) = \sigma^2 + \nu^2,$$
$$S(W) = 2\left(1 + \frac{\sigma^2}{\nu^2}\right)^{-3/2} \quad \text{and} \quad K(W) = 6\left(1 + \frac{\sigma^2}{\nu^2}\right)^{-2},$$

respectively.

Based on the *odd log-logistic generator* (OLL-G) class [6], we define the OLLExGa cdf, say $F(y) = F(y; \mu, \sigma, \nu, \tau)$, by integrating the log-logistic density function with shape parameter $\tau > 0$, namely

$$F(y) = \int_0^{G_{\mu,\sigma,\nu}(y)/\bar{G}_{\mu,\sigma\nu}(y)} \frac{\tau x^{\tau-1}}{(1 + x^\tau)^2} \mathrm{d}x = \frac{G_{\mu,\sigma,\nu}(y)^\tau}{G_{\mu,\sigma,\nu}(y)^\tau + \bar{G}_{\mu,\sigma,\nu}(y)^\tau}, \tag{3}$$

where $\bar{G}_{\mu,\sigma,\nu}(y) = 1 - G_{\mu,\sigma,\nu}(y)$. Hereafter, we assume that the random variable $Y$ follows the cdf (3) with parameters $(\mu, \sigma, \nu, \tau)^{\mathrm{T}}$, say $Y \sim \mathrm{OLLExGa}(\mu, \sigma, \nu, \tau)$. The OLLExGa distribution includes as special cases the ExGa distribution when $\tau = 1$ and the normal distribution when $\tau = 1$ and $\nu = 0$.

Consider $\eta(y) = G_{\mu,\sigma,\nu}(y)$ to simplify the notation. The density function of $Y$ has the form

$$\begin{aligned} f(y) = f(y; \mu, \sigma, \nu, \tau) = \frac{\tau}{\nu} \exp\left(\frac{\mu - y}{\nu} + \frac{\sigma^2}{2\nu^2}\right) \Phi\left(\frac{y - \mu}{\sigma} - \frac{\sigma}{\nu}\right) \\ \times \left\{\eta(y)[1 - \eta(y)]\right\}^{\tau-1} \left\{\eta(y)^\tau + [1 - \eta(y)]^\tau\right\}^{-2}. \end{aligned} \tag{4}$$

The main motivation for the new distribution is to make its skewness more flexible (compared to the ExGa model) and allow bimodality. Equation (4) provides greater flexibility of the tails of the density and can be widely applied in many areas of engineering and biology.

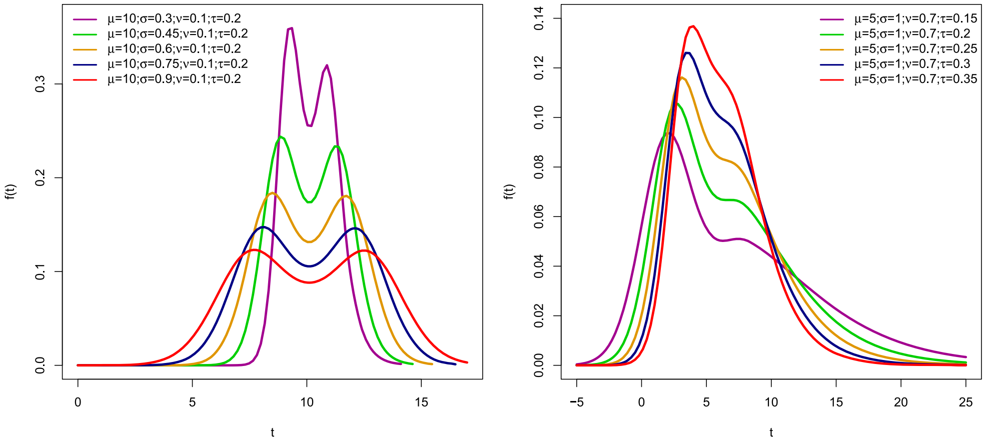
Plots of the density (4) for selected parameter values are displayed in Figure 2. It is clear that the proposed distribution is much more flexible, especially in relation to bimodality (for $0 < \tau < 0.5$) than the ExGa distribution, which does not have this characteristic.

The quantile function (qf) of the OLLExGa distribution can be expressed as

$$y = Q_{\mathrm{ExGa}}\left(\frac{u^{1/\tau}}{u^{1/\tau} + [1 - u]^{1/\tau}}\right), \tag{5}$$

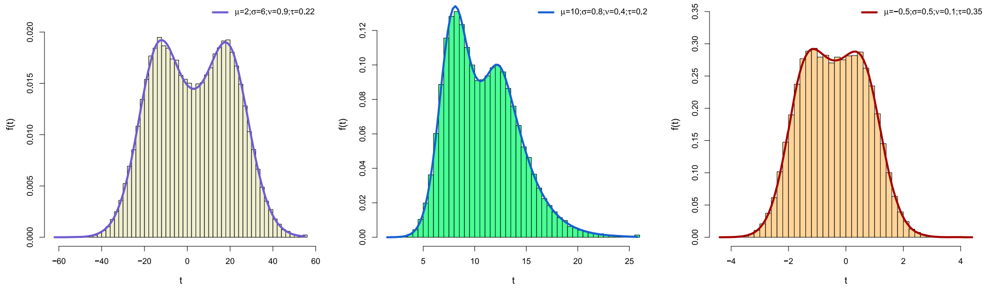
where $Q_{\mathrm{ExGa}}(u) = G^{-1}_{\mu,\sigma,\nu}(u)$ is the qf of the ExGa distribution available in the GAMLSS package [27].

This scheme is useful because of the existence of fast generators for the ExGa random variables in some statistical packages. The plots comparing the exact OLLExGa densities and the histograms from two simulated data sets with 100,000 replications for selected parameter values are displayed in Figure 3. These plots (and several others not shown here) reveal that the simulated values are consistent with the OLLExGa distribution.



(a) (b)

**Figure 2.** Plots of the OLLExGa density for some parameter values.



(a) (b) (c)

**Figure 3.** Histograms and plots of the OLLExGa densities.

In the Appendix, we derive some mathematical properties of the OLLExGa distribution including a linear representation for its density function.

## 3. The OLLExGa regression

In several problems of the medical, biological, industrial and chemical areas, among others, it is of great interest to verify if two or more variables are related in some way. To investigate this relationship is very important to construct a regression model. The data collection allows to know the nature of the relationship between variables and to carry out studies capable of accommodating unexpected situations, such as variability in raw material, ambient temperature, machine and operators. They are built with the following objectives: model formulation, parameter estimation, inference, diagnostic and residual analysis and prediction. In this research, we focus on our these goals. In these terms, the OLLExGa regression is a very competitive alternative to the ExGa regression.

The regression technique aims to choose the distribution of $Y$ given the matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ of explanatory variables. The parameters $\mu$ and $\sigma$ are related to the explanatory variables by the systematic components

$$\mu_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_1 \quad \text{and} \quad e\sigma_i = \exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_2), \quad i = 1, \ldots, n, \tag{6}$$

respectively, where $\mathbf{x}_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{ip})$ and $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p})^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2p})^{\mathrm{T}}$ are the unknown vectors of coefficients.

The total log-likelihood function for the vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \nu, \tau)^{\mathrm{T}}$ from model (6) given $n$ independent observations $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ has the form

$$\begin{aligned} l(\boldsymbol{\theta}) = {} & n \log(\tau) - n \log(\nu) + \sum_{i=1}^{n} \left( \frac{\mu_i - y_i}{\nu} + \frac{\sigma_i^2}{2\nu^2} \right) + \sum_{i=1}^{n} \log \Phi \left( \frac{y_i - \mu_i}{\sigma_i} + \frac{\sigma_i}{\nu} \right) \\ & + (\tau - 1) \sum_{i=1}^{n} \log\{\eta(y_i)[1 - \eta(y_i)]\} - 2 \sum_{i=1}^{n} \log\{\eta(y_i)^{\tau} + [1 - \eta(y_i)]^{\tau}\}. \end{aligned} \tag{7}$$

The log-likelihood (7) can be maximized numerically using the GAMLSS software to find the MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. By fitting the ExGa regression (with $\tau = 1$) yields initial values for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Some simulations of the fitted model (6) confirm the adequacy of this maximization in Section 3.1.

The elements of the $(2p + 2) \times (2p + 2)$ Hessian matrix $\ddot{\mathbf{L}}(\boldsymbol{\theta})$ can be determined numerically in the **R** software. The multivariate normal distribution $N_{2p+2}(0, -\ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}})^{-1})$ can approximate the distribution of $\widehat{\boldsymbol{\theta}}$ since Equation (4) satisfies some standard regularity conditions. More importantly, it can be utilized to obtain approximate confidence intervals for the parameters in $\boldsymbol{\theta}$. The adequacy of some special models of the OLLExGa regression can be verified via likelihood ratio (LR) statistics.

### *3.1. Two simulation studies*

In this section, we provide two simulation studies: one to examine the adequacy of the MLEs in the OLLExGa distribution and other to investigate the adequacy of the estimates in the regression model with systematic components for $\mu$ and $\sigma$.

- *First simulation: the OLLExGa distribution*
  Some properties of the MLEs are examined using a classical analysis by means of a Monte Carlo simulation study. We simulate the OLLExGa distribution as follows: (i) Generate $u \sim U(0, 1)$; (ii) Obtain OLLExGa observations $y = Q_{ExGa}(u)$ from Equation (5).
  We set $\mu = 10, \sigma = 0.8, \nu = 0.4$ and $\tau = 0.2$ to provide bimodality in the data as shown in Figure 3(b). We choose four scenarios ($n$ = 50, 100, 500 and 1000) for the replications to calculate $\hat{\mu}, \hat{\sigma}, \hat{\nu}$ and $\hat{\tau}$. Then, we obtain the average estimates (AEs), biases and means square errors (MSEs) from 1000 Monte Carlo simulations via the GAMLSS software. The results listed in Table 1 confirm the accuracy of the estimates and that their MSEs decrease when $n$ increases in agreement with first-order asymptotic theory.
- *Second Simulation: the OLLExGa regression*
  We examine the performance of the MLEs in the OLLExGa regression by means of some

**Table 1.** AEs, biases and MSEs for the parameters of the OLLExGa distribution.

| | Scenario 1 | | | | Scenario 2 | | |
|---|---|---|---|---|---|---|---|
| | $n = 50$ | | | | $n = 100$ | | |
| Parameter | AE | Bias | MSE | Parameter | AE | Bias | MSE |
| $\mu$ | 9.1779 | −0.8221 | 10.7552 | $\mu$ | 9.4233 | −0.5767 | 4.2744 |
| $\sigma$ | 1.1141 | 0.3141 | 0.3622 | $\sigma$ | 0.9858 | 0.1858 | 0.1772 |
| $\nu$ | 0.5415 | 0.1415 | 1.4680 | $\nu$ | 0.6357 | 0.2357 | 0.5249 |
| $\tau$ | 0.4363 | 0.2363 | 0.2081 | $\tau$ | 0.3335 | 0.1335 | 0.1446 |
| | Scenario 3 | | | | Scenario 4 | | |
| | $n = 500$ | | | | $n = 1000$ | | |
| Parameter | AE | Bias | MSE | Parameter | AE | Bias | MSE |
| $\mu$ | 9.8203 | −0.1797 | 1.1707 | $\mu$ | 9.8306 | −0.1694 | 0.5779 |
| $\sigma$ | 0.9151 | 0.1151 | 0.0376 | $\sigma$ | 0.8952 | 0.0952 | 0.0214 |
| $\nu$ | 0.5886 | 0.1886 | 0.1332 | $\nu$ | 0.5734 | 0.1734 | 0.0750 |
| $\tau$ | 0.2508 | 0.0508 | 0.0083 | $\tau$ | 0.2391 | 0.0391 | 0.0038 |

simulations with $n = 100$, 300 and 500. We simulate 1,000 samples from two scenarios ($\tau = 0.5$ and $\tau = 1.3$). For both cases, we take $\beta_{10} = 2.1$, $\beta_{11} = -0.4$, $\beta_{12} = 0.3$, $\beta_{20} = -1$, $\beta_{21} = 0.2$, $\beta_{22} = -0.1$ and $\nu = 0.4$ under the systematic components $\mu_i = \beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2}$ and $\sigma_i = \beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2}$. The response variable $Y_i$ and explanatory variables $X_{i1}$ and $X_{i2}$ are generated as follows: $Y_i \sim \text{OLLExGa}\ (\mu_i, \sigma_i, \nu, \tau)$, $X_{i1} \sim \text{Uniform}\ (0, 1)$ and $X_{i2} \sim \text{Binomial}\ (2, 0.5)$.

We calculate the AEs, biases and MSEs for each fitted regression. The figures in Table 2 reveal that the MSEs of the estimates tend to zero and the AEs converge to the true

**Table 2.** AEs, biases and MSEs for the OLLExGa regression under scenarios 1 ($\tau = 0.5$) and 2 ($\tau = 1.3$).

| | Scenario 1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $n = 100$ | | | $n = 300$ | | | $n = 500$ | | |
| Parameter | AE | Bias | MSE | AE | Bias | MSE | AE | Bias | MSE |
| $\beta_{10}$ | 2.0729 | −0.0271 | 0.1386 | 2.0861 | −0.0139 | 0.0537 | 2.0793 | −0.0207 | 0.0332 |
| $\beta_{11}$ | −0.4115 | −0.0115 | 0.1359 | −0.4128 | −0.0128 | 0.0442 | −0.4007 | −0.0007 | 0.0262 |
| $\beta_{12}$ | 0.3035 | 0.0035 | 0.0251 | 0.3011 | 0.0011 | 0.0076 | 0.3031 | 0.0031 | 0.0046 |
| $\beta_{20}$ | −1.1173 | −0.1173 | 0.2861 | −1.0281 | −0.0281 | 0.0807 | −1.0163 | −0.0163 | 0.0447 |
| $\beta_{21}$ | 0.2438 | 0.0438 | 0.4549 | 0.1978 | −0.0022 | 0.0971 | 0.2107 | 0.0107 | 0.0525 |
| $\beta_{22}$ | −0.1085 | −0.0085 | 0.0706 | −0.1101 | −0.0101 | 0.0146 | −0.0998 | 0.0002 | 0.0068 |
| $\nu$ | 0.4499 | 0.0499 | 0.1129 | 0.4282 | 0.0282 | 0.0497 | 0.4254 | 0.0254 | 0.0329 |
| $\tau$ | 0.5468 | 0.0468 | 0.0948 | 0.5225 | 0.0225 | 0.0407 | 0.5225 | 0.0225 | 0.0282 |
| | Scenario 2 | | | | | | | | |
| | $n = 100$ | | | $n = 300$ | | | $n = 500$ | | |
| Parameter | AE | Bias | MSE | AE | Bias | MSE | AE | Bias | MSE |
| $\beta_{10}$ | 2.1881 | 0.0881 | 0.0534 | 2.1125 | 0.0125 | 0.0259 | 2.1021 | 0.0021 | 0.0186 |
| $\beta_{11}$ | −0.4042 | −0.0042 | 0.0216 | −0.4008 | −0.0008 | 0.0065 | −0.3979 | 0.0020 | 0.0037 |
| $\beta_{12}$ | 0.2997 | −0.0003 | 0.0034 | 0.3022 | 0.0022 | 0.0011 | 0.2999 | −0.0001 | 0.0007 |
| $\beta_{20}$ | −1.1240 | −0.1240 | 0.7935 | −1.0332 | −0.0332 | 0.2199 | −1.0255 | −0.0255 | 0.0919 |
| $\beta_{21}$ | 0.2087 | 0.0087 | 0.9323 | 0.2093 | 0.0093 | 0.0976 | 0.2032 | 0.0032 | 0.0417 |
| $\beta_{22}$ | −0.1009 | −0.0009 | 0.0713 | −0.1004 | −0.0004 | 0.0150 | −0.1038 | −0.0038 | 0.0079 |
| $\nu$ | 0.3060 | −0.0940 | 0.0677 | 0.3875 | −0.0125 | 0.0359 | 0.4001 | 0.0001 | 0.0270 |
| $\tau$ | 1.4524 | 0.1524 | 1.5819 | 1.4080 | 0.1080 | 0.9306 | 1.3395 | 0.0395 | 0.4188 |

parameters when $n$ increases. Both facts strongly support that the approximate normal distribution is adequate to the finite sample distribution of the estimates.

## 4. Checking model

The assessment of robustness aspects of the parameter estimates in statistical models has been an important concern of various researchers in recent decades. The case deletion measures, which consists of studying the impact on the parameter estimates after dropping individual observations, is probably the most employed technique to detect influential observations; see, for example [3]. A global influence measure considered by [30] is a generalization of the Cook distance defined as a standardized norm of $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$ expressed as

$$GD_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^\top \left[-\ddot{\mathbf{L}}(\boldsymbol{\theta})\right] (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}), \tag{8}$$

where $-\ddot{\mathbf{L}}(\boldsymbol{\theta})$ is the observed information matrix. Another measure to evaluate the influence is called of likelihood distance and considers the difference between $\hat{\boldsymbol{\theta}}_{(i)}$ and $\hat{\boldsymbol{\theta}}$. Thus, the likelihood distance has the form

$$LD_i(\boldsymbol{\theta}) = 2\left[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)})\right], \tag{9}$$

where $l(\hat{\boldsymbol{\theta}})$ is the value of the logarithm of the likelihood function of the full sample and $l(\hat{\boldsymbol{\theta}}_{(i)})$ is the value of the logarithm of the likelihood function of the sample excluding the $i$th observation.

The analysis of the residuals is an efficient method to check the model adequacy. Recently, [23] presents a discussion and application of the qrs for regression models. Here, we also use the qrs to check the adequacy of the OLLExGa regression. The residuals usually allow to check the local fit to each observation and whether the differences between the observed and fitted values occur randomly or are due to a systematic behavior. The qrs [5] for model (6) are defined by

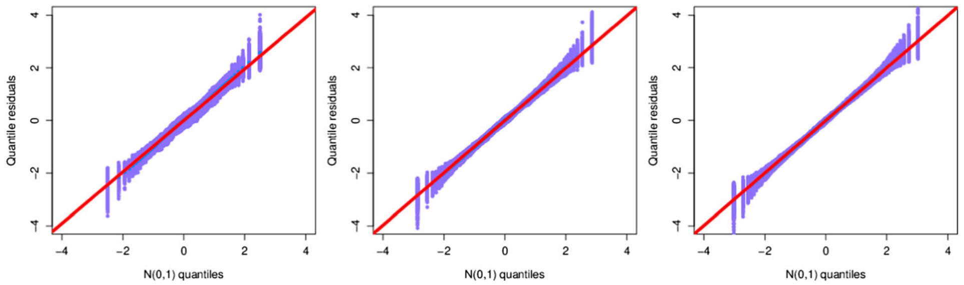
$$qr_i = \Phi^{-1}\left\{\frac{\eta(y_i)^\tau}{\eta(y_i)^\tau + [1-\eta(y_i)]^\tau}\right\}, \tag{10}$$

where $\eta(y) = G_{\mu,\sigma,\nu}(y)$ and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution.

The construction of simulated confidence bands to provide a better interpretation of the probability normal plot of the residuals was pioneered by [1]. The majority of points will be randomly distributed within these bands when the model is well-suited to the data.

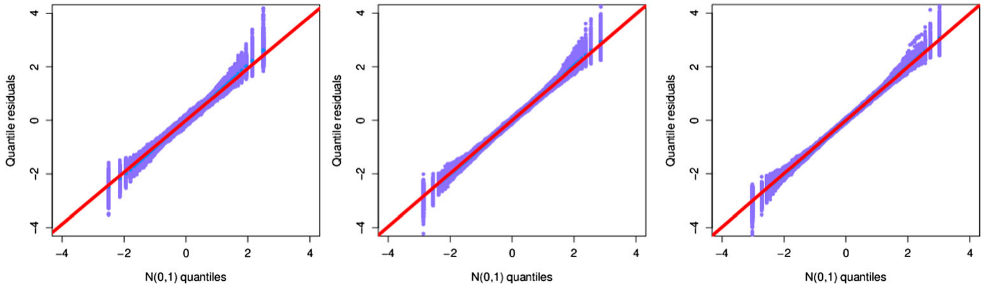
*Simulation study of the quantile residuals*

The behavior of the empirical distribution of the $qr_i's$ for the OLLExGa regression is investigated by generating 1, 000 samples via the algorithm introduced in Section 3.1. We construct the normal probability plot to check the deviation from the normality hypothesis for the residuals. The plots in Figures 4 and 5 representing the first and second scenarios, respectively, indicate that the empirical distribution of these residuals agrees with the standard normal distribution. Also, this empirical distribution becomes closer to the standard normal distribution when $n$ increases



(a) (b) (c)

**Figure 4.** Normal probability plots for $qr_i's$ in the OLLExGa regression for scenario 1 ($\tau = 0.5$) (a) $n = 100$. (b) $n = 300$. (c) $n = 500$.



(a) (b) (c)

**Figure 5.** Normal probability plots for $qr_i's$ in the OLLExGa regression for scenario 2 ($\tau = 1.3$) (a) $n = 100$. (b) $n = 300$. (c) $n = 500$.

## 5. Applications

In this section, we present three real applications in the field of agriculture, where we prove that the OLLExGa regression can be quite useful in this area. The calculations are performed with the **R** software.

### *5.1. Application 1: tomato seeds data*

The data refer to index of germination speed of tomato seeds *Ozone*. The production of tomatoes for fresh consumption generally involves the germination of seeds in trays with subsequent transplanting of the seedlings. One of the main problems noted in this production system is the rapid vegetative growth of the aerial part (etiolation). This imbalance causes the formation of elongated and fragile seedlings with thin hypocotyls and few roots, making them more susceptible to biotic and abiotic stresses, with consequent death of these seedlings [25]. Some growth inhibitors, such as paclobutrazol (PBZ), are used to reduce this problem. PBZ is a growth regulator that belongs to the triazole group and acts by reducing biosynthesis of gibberellins (GAs). It therefore reduces the growth of the stem without impairing cell differentiation and without causing phytotoxicity [17]. The gibberellins are hormones responsible for regulating the height of plants, by promoting alteration of the juvenility and sexuality of the flowers and the establishment and growth of the fruits, besides affecting the activation of hydrolytic enzymes responsible for seed germination [28]. PBZ can be applied by foliar spraying, by soil applications or by seed treatment. The application by seed treatment is one of the safest options by avoiding the problem of residues in the fruits and environmental contamination [15]. However, since PBZ acts by reducing the synthesis of GAs, it can have deleterious effects on seed germination. Calculation of the germination speed index (GSI) proposed by [16] can be used as a test of the relative vigor of seeds in controlled laboratory germination experiments. There is a direct relationship between the germination sped and vigor of seeds [19]. Therefore, a reduction of the GSI serves as an indicator of a negative influence on seed germination.

*Method for testing the seed germination index of tomatoes*

The study was carried out in the Central Seed Laboratory of Lavras Federal University (UFLA), in the municipality of Lavras, Minas Gerais, Brazil. Seeds of the *Ozone* tomato cultivar were treated with four PBZ doses 0, 0.004, 0.008 and 0.016 mL/10 g of seeds, with the dose of 0.008 recommended by the manufacturer. So that all treatments receive the same volume of solution (0.214 mL), complementary water was applied. The volume of the solution was distributed as uniformly as possible in Petri dishes with a diameter of 25 cm, after which the seeds were added and the dishes were covered with lids and manually shaken for approximately 4 min. After the treatments, a portion of the seeds from each dish was submitted to analysis (period 1) and the other portion was placed in a paper bag and stored in a refrigerator (10°C, 50% RH) for 5 months (period 2). The GSI was calculated daily by counting the number of germinated seeds (with radicle emergence) in the germination tests, using acrylic gerboxes containing blotter paper as substrate, moistened to 2.5 times the dry weight. The gerboxes were kept in BOD chambers at temperature of 20°–30°C for 14 days, with 12:12 h photoperiod. The GSI values were estimated using the formula proposed by [16], namely

$$GSI = \frac{G_1}{N_1} + \cdots + \frac{G_n}{N_n}, \tag{11}$$

where

- $G_1, G_2, \ldots, G_n$ denote the number of normal seedlings tallied on the first, second, ..., last count, respectively;
- $N_1, N_2, \ldots, N_n$ denote the number of days since sowing on the first, second, ..., last count, respectively.

Thus, we adopt a regression based on the OLLExGa distribution to model these data considering two systematic components for $\mu$ and $\sigma$. The objective is to verify if any of the doses in different periods is relevant in determining the GSI. The variables under study are:

- $y_i$: germination speed index (GSI);
- $x_{i1}$: doses of PBZ (0, 0.004, 0.008 and 0.016 mL) per 10 grams of seeds. In this case using three dummy variables;
- $x_{i2}$: two periods (0 = period 1, 1 = period 2).

**Table 3.** Descriptive statistics for the GSI response variable.

| Mean | Median | SD | Skewness | Kurtosis | Min. | Max. |
|---|---|---|---|---|---|---|
| 11.0020 | 8.5650 | 5.1852 | 0.5635 | −0.2438 | 3.9100 | 27.2500 |

Table 3 provides the descriptive analysis of the response variable, where the mean and median are 11.0020 and 8.5650, respectively. The distribution of the data is asymmetric and then the OLLExGa distribution is an alternative to the analysis of the data.

In Table 4, we give the MLEs, their standard errors (SEs) (in parentheses) and some goodness-of-fit measures: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Global Deviance (GD) for three fitted distributions. The figures in this table indicate that the OLLExGa distribution has the lowest values of these statistics. In fact, it is the best model to fit the current data.

The OLLExGa distribution includes as a special case the ExGa model, and then we can compare them using the LR statistic. This statistic for testing the hypotheses $H_0 : \tau = 1$ versus $H_1 : \tau \neq 1$, i.e. to compare the OLLExGa and ExGa distributions, is $w = 20.8879$ ($p-$value $=< 0.0001$). It is evident that the OLLExGa distribution provides a better fit to these data than the ExGa distribution.

More information is addressed by a visual comparison of the histogram of the data and the estimated densities and cumulative functions. The plots of the fitted OLLExGa, ExGa and normal densities are displayed in Figure 6. The estimated OLLExGa density gives the closest fit to this histogram.

In addition, we note that the tomato seeds data have a bimodal shape in Figure 6, whereas the ExGa and normal distributions cannot cope with this shape. So, this plot indicates that the OLLExGa regression is a possible model to explain the current data. Based on this marginal and descriptive analysis, we consider the OLLExGa regression with two systematic components:

$$\begin{aligned}\mu_i = \beta_{10} + \beta_{111}x_{i11} + \beta_{112}x_{i12} + \beta_{113}x_{i13} + \beta_{12}x_{i21} \\ + \beta_{131}(x_{i11} \times x_{i21}) + \beta_{132}(x_{i12} \times x_{i21}) + \beta_{133}(x_{i13} \times x_{i21})\end{aligned}$$

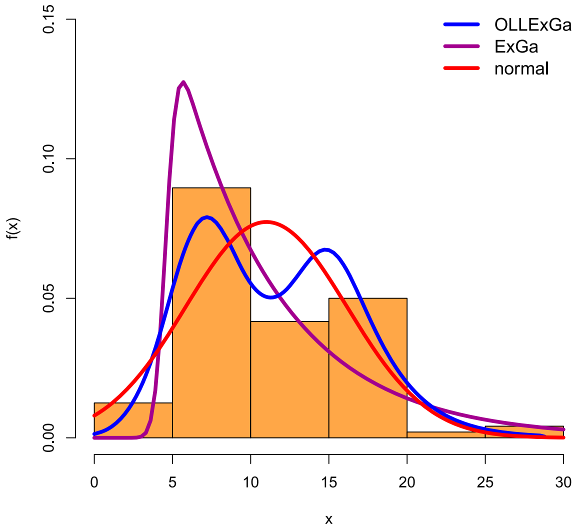
and

$$\begin{aligned}\sigma_i = \exp\{\beta_{20} + \beta_{211}x_{i11} + \beta_{212}x_{i12} + \beta_{213}x_{i13} + \beta_{22}x_{i21} \\ + \beta_{231}(x_{i11} \times x_{i21}) + \beta_{232}(x_{i12} \times x_{i21}) + \beta_{233}(x_{i13} \times x_{i21})\}, \quad i = 1, \ldots, 96,\end{aligned}$$

where $(x_{i11} \times x_{i21})$, $(x_{i12} \times x_{i21})$ and $(x_{i13} \times x_{i21})$ refer to the dose–period interaction.

**Table 4.** MLEs, SEs and AIC, BIC and GD statistics for some models fitted to the tomato seeds data.

| Model | $\mu$ | $\log(\sigma)$ | $\log(\nu)$ | $\tau$ | AIC | BIC | GD |
|---|---|---|---|---|---|---|---|
| OLLExGa | 10.7499 | 0.2458 | −0.8386 | 0.1690 | 551.0679 | 561.3253 | 543.0679 |
| | (0.0057) | (0.1718) | (0.0002) | (0.0373) | | | |
| ExGa | 4.5912 | −0.5605 | 1.8585 | 1 | 569.9559 | 577.6489 | 563.9559 |
| | (0.2914) | (0.4579) | (0.1114) | (–) | | | |
| normal | 11.0022 | 1.6405 | (–) | (–) | 591.4290 | 596.5577 | 587.4290 |
| | (0.5265) | (0.0722) | (–) | (–) | | | |



(a)

**Figure 6.** Estimated densities of the OLLExGa, ExGa and normal models for tomato seeds data.

**Table 5.** MLEs, SEs and *p*-values for the OLLExGa regression fitted to the tomato seeds data.

| | Sources of variation | Parameter | Estimate | SE | *p*-value |
|---|---|---|---|---|---|
| | Intercept | $\beta_{10}$ | 6.3362 | 2.2801 | 0.0068 |
| | Dose 0.004 | $\beta_{111}$ | −0.8545 | 0.4373 | 0.0542 |
| | Dose 0.008 | $\beta_{112}$ | −1.2836 | 0.3994 | 0.0019 |
| $\mu$ | Dose 0.016 | $\hat{\beta}_{113}$ | 0.4649 | 0.3612 | 0.2018 |
| | Period 2 | $\hat{\beta}_{12}$ | 9.7536 | 2.3479 | $< 0.001$ |
| | Dose 0.004 × Period 2 | $\beta_{131}$ | −0.3778 | 2.3813 | 0.8744 |
| | Dose 0.008 × Period 2 | $\beta_{132}$ | −0.1861 | 0.9406 | 0.8437 |
| | Dose 0.016 × Period 2 | $\beta_{133}$ | −1.2384 | 0.9121 | 0.1785 |
| | Intercept | $\beta_{20}$ | 1.8998 | 0.9169 | 0.0416 |
| | Dose 0.004 | $\beta_{211}$ | 0.0011 | 0.0106 | 0.9160 |
| | Dose 0.008 | $\beta_{212}$ | −0.1719 | 0.2842 | 0.5470 |
| $\sigma$ | Dose 0.016 | $\hat{\beta}_{213}$ | −0.4708 | 0.2902 | 0.1088 |
| | Period 2 | $\hat{\beta}_{22}$ | 0.7969 | 0.3192 | 0.0147 |
| | Dose 0.004 × Period 2 | $\beta_{231}$ | −0.7715 | 0.3595 | 0.0350 |
| | Dose 0.008 × Period 2 | $\beta_{232}$ | 0.1658 | 0.4678 | 0.7239 |
| | Dose 0.016 × Period 2 | $\beta_{233}$ | 0.4448 | 0.4728 | 0.3497 |
| | | $\log(\nu)$ | −0.8481 | 5.3061 | |
| | | $\tau$ | 6.9780 | 6.2670 | |

Table 5 gives the MLEs, their approximate SEs and $p$-values obtained from the fitted OLLExGa regression. We can note from Table 5 that the covariate $x_1$ is significant considering a level of 5%, which indicates that there is a significant difference between the levels of the doses 0.000 and 0.008. In the systematic component referring to the dispersion parameter ($\sigma$), the factor $x_2$ is significant, and then there is a significant difference between times 1 and 2. Other interpretations are reported at the end of this application.

Table 6 lists the AIC, BIC and GD statistics for some fitted regressions. The results indicate that the OLLExGa regression has the smallest values of these statistics among all of

**Table 6.** AIC, BIC and GD statistics for some fitted regressions to the tomato seeds data.

| Model | AIC | BIC | GD |
|---|---|---|---|
| OLLExGa | 370.4092 | 416.5674 | 334.4092 |
| normal | 376.1662 | 417.1958 | 344.1662 |
| ExGa | 378.2402 | 421.8341 | 344.2402 |

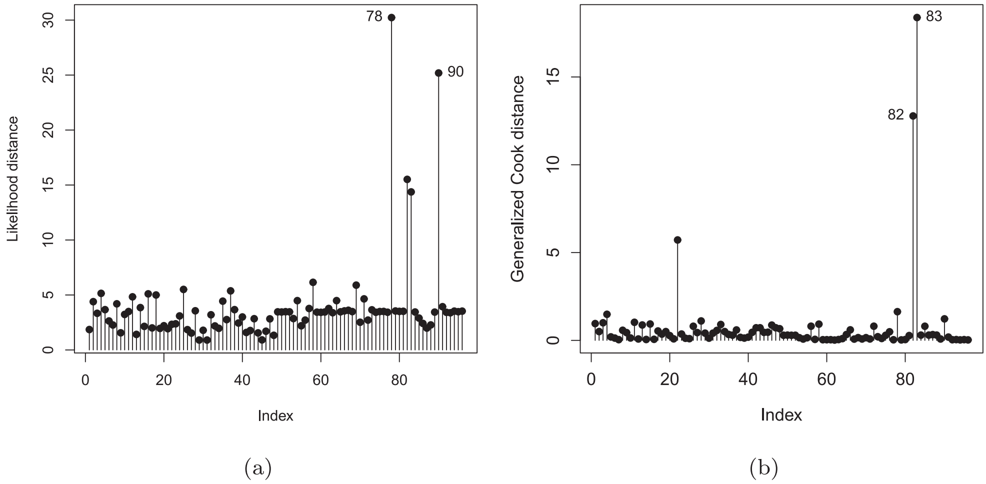


**Figure 7.** Index plot for $\boldsymbol{\theta}$: (a) $LD_i(\boldsymbol{\theta})$ (likelihood distance) and (b) $GD_i(\boldsymbol{\theta})$ (generalized Cook's distance).

them. So, it could be chosen as the more suitable regression to these data. The LR statistic for testing the hypotheses $H_0 : \tau = 1$ versus $H_1 : \tau \neq 1$, i.e. to compare the OLLExGa and ExGa regressions, is $w = 9.8310$ ($p$—value $= 0.0017$), which supports the OLLExGa regression.

We use the software **R** to compute case deletion measures $LD_i(\boldsymbol{\theta})$ and $GD_i(\boldsymbol{\theta})$ defined in Section 4. The results of such influence measure index plots are displayed in Figure 7. These plots show that the cases ♯78, ♯82, ♯83 and ♯90 are possible influential observations.

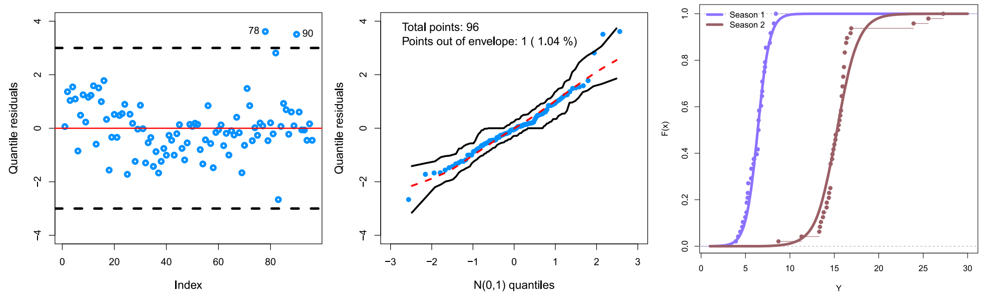
We perform the residual analysis by plotting in Figure 8(a) the $qr_i$'s (see Section 4) against the index of the observations. Figure 8(b) gives the normal probability plot with generated envelope. Figure 8(a) shows some large residuals (observations ♯78 and ♯90), although Figure 8(b) supports the hypothesis that the OLLExGa regression is very suitable for these data.

After removing some non-significant explanatory variables, the final model has the form

$$\mu_i = \beta_{10} + \beta_{111}x_{i11} + \beta_{112}x_{i12} + \beta_{113}x_{i13} + \beta_{12}x_{i21}$$

and

$$\sigma_i = \exp\{\beta_{20} + \beta_{22}x_{i21} + \beta_{231}(x_{i11} \times x_{i20}) + \beta_{232}(x_{i11} \times x_{i21}) + \beta_{233}(x_{i12} \times x_{i20}) \\ + \beta_{234}(x_{i12} \times x_{i21}) + \beta_{235}(x_{i13} \times x_{i20}) + \beta_{236}(x_{i13} \times x_{i21})\}.$$



(a) (b) (c)

**Figure 8.** (a) Index plot of the qrs ($qr_i$). (b) Normal probability plot with envelope for $qr_i$. (c) Estimated cdf from the OLLExGa regression for the tomato seeds and the empirical cdf for levels of the variable $x_2$.

**Table 7.** MLEs, SEs and $p$-values for the final OLLExGa regression fitted to the tomato seeds data.

| | Sources of variation | Parameter | Estimates | SE | $p$-value |
|---|---|---|---|---|---|
| $\mu$ | Intercept | $\beta_{10}$ | 5.9078 | 0.8288 | $< 0.001$ |
| | Dose 0.004 | $\beta_{111}$ | −0.8002 | 0.3649 | 0.0311 |
| | Dose 0.008 | $\beta_{112}$ | −1.3167 | 0.3610 | 0.0005 |
| | Dose 0.016 | $\beta_{113}$ | 0.4241 | 0.3434 | 0.2204 |
| | Period 2 | $\beta_{12}$ | 9.2285 | 0.3235 | $< 0.001$ |
| $\sigma$ | Intercept | $\beta_{20}$ | 2.0641 | 1.4065 | 0.1461 |
| | Period 2 | $\beta_{22}$ | 0.8026 | 0.3664 | 0.0314 |
| | Dose 0.004 × Period 1 | $\beta_{231}$ | 0.0088 | 0.3319 | 0.9789 |
| | Dose 0.004 × Period 2 | $\beta_{232}$ | −0.7739 | 0.3632 | 0.0361 |
| | Dose 0.008 × Period 1 | $\beta_{233}$ | −0.1735 | 0.3386 | 0.6098 |
| | Dose 0.008 × Period 2 | $\beta_{234}$ | −0.0075 | 0.3781 | 0.9841 |
| | Dose 0.016 × Period 1 | $\beta_{235}$ | −0.4781 | 0.3408 | 0.1645 |
| | Dose 0.016 × Period 2 | $\beta_{236}$ | 0.0149 | 0.3879 | 0.9695 |
| | | $\log(\nu)$ | −0.1194 | 0.8552 | |
| | | $\tau$ | 8.3140 | 11.3920 | |

### *Some interpretations for the final regression*

- The numbers in Table 7 indicate that the covariable variables $x_1$ (representing dose) is significant at 5%, meaning a significant difference between the doses 0.000 vs. 0.004 and 0.000 vs. 0.008 in relation to the GSI. The dose of 0.008 (mL i.e. 10 g seeds$^{-1}$) is the one recommended by the manufacturer of the product (Syngenta) for tomatoes. Besides this, it has not yet been demonstrated that this dose is sufficient for control of etiolation, because even though this dose is recommended, it is very low (32$\mu$l commercial product (Pc) 10 g seeds$^{-1}$) in relation to the volume of seeds. Tomato seeds are very small (1000 seeds weigh about 3 to 3.5 g), making it hard to perform uniform seed treatment. For this reason, we tested twice this dose and by interpolation and intermediate dose (0.012 mL i.a. 10 g seeds$^{-1}$). But as observed in these results, the product reduces the physiological quality of the seeds even when applied in the recommended dose. In agreement with the results found in the conventional statistical analyzes [24].
- We also observed that the covariable $x_2$ is significant (5%), indicating the existence of a significant difference between periods 1 and 2 in relation to the GSI. It was observed that GSI at all doses was higher at period 2 (Figure 8(c)), and this may be due to a lower

reactivity of the product at period 2 associated with the biological factor (seed deterioration). The residual period (time when the chemical or biological product's active ingredient continues to be effective in the environment where it is used) of the PBZ is 180 days. Since the seeds were treated and stored for 150 days (period 2), the potential harmful effect of the product on seed germination was reduced. In addition, a higher GSI in period 2 compared to 1 may be due to a longer delay in structuring the seed membrane systems at the moment of germination, allowing a faster imbibition (absorption of a liquid by a solid), with consequent faster germination speed. In this case, the larger GSI does not necessarily mean better quality. Confirming this would require associating this result with others obtained by the same method in other tests (seed germination, seedling emergence, among others).

- Note that the interaction of the covariables ($x_1 \times x_2$) is significant, meaning a relevant difference exists of the interaction of the dose of 0.004 mL in period 2. This result is contrary to expectation, since this dose is half that recommended by the manufacturer. Therefore, our expectation was that there would be no damaging effect on the seed vigor even in period 1, when the product was fully active. In contrast, in period 2 we observed lower activity of the product.

Finally, a graphical comparison between the levels of the variable $x_2$ is illustrated in Figure 8(c). These plots refer to the empirical cdf and the estimated cdf of the OLLExGa regression. They confirm that the OLLExGa regression provides a superior fit. Also, we can note a significant difference between the epochs in relation to the speed index of germination of seeds of tomato cultivar Ozone.

### 5.2. Application 2: weight of rat pups data

For the second application, we consider only the regression structure for the parameter $\mu$. The data refer to the birth weights of rats, see [20]. The sample size is $n = 322$ and we consider the following variables:

- $y_i$: weight of newborn rats;
- $x_{i1}$: sex ($0 =$ female, $1 =$ male);
- $x_{i2}$: treatment ($0 =$ control, $1 =$ low, $2 =$ high). In this case, we take two dummy variables ($i = 1, \ldots, 322$).

A brief descriptive analysis of the data in Table 8 reveals that the mean weight is 6.081 g and the median is 6.055 g, thus indicating that the data have symmetrical shape. Because of this fact, we can compare the OLLExGa, ExGa and normal regression models. The OLLExGa regression is indeed flexible and it is considered not only for bimodal asymmetric data, but also for symmetric data.

**Table 8.** Descriptive statistics for the weights of rat pups data.

| Mean | Median | SD | Skewness | Kurtosis | Min. | Max. |
|---|---|---|---|---|---|---|
| 6.081 | 6.055 | 0.6474 | 0.4970 | 1.1251 | 3.680 | 8.330 |

**Table 9.** MLEs, SEs and $p$-values for the fitted OLLExGa regression to the weights of rat pups data.

| Parameter | Estimate | SE | $p$-value |
|---|---|---|---|
| $\beta_{10}$ | 5.9099 | 0.0611 | $< 0.001$ |
| $\beta_{11}$ | 0.2023 | 0.0653 | 0.0021 |
| $\beta_{12}$ | −0.4608 | 0.0789 | $< 0.001$ |
| $\beta_{13}$ | −0.3852 | 0.0785 | $< 0.001$ |
| $\log(\sigma)$ | 0.9236 | 0.0431 | |
| $\log(\nu)$ | −1.2171 | 0.1093 | |
| $\tau$ | 4.8450 | 0.2096 | |

**Table 10.** Goodness-of-fit measures for the weights of rat pups data.

| Model | AIC | BIC | GD |
|---|---|---|---|
| OLLExGa | 585.4798 | 611.9016 | 571.4798 |
| ExGa | 593.1629 | 615.8103 | 581.1629 |
| normal | 598.2369 | 617.1097 | 588.2369 |

The OLLExGa regression with only a systematic component for $\mu$ is

$$\mu_i = \beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2(0-1)} + \beta_{13}x_{i2(0-2)}, \quad i = 1, \ldots, 322.$$

Table 9 provides the MLEs, their approximate SEs and $p$-values obtained from the fitted OLLExGa regression. We conclude that the two explanatory variables are significant at a 5% significant level. Thus, we can confirm (under this risk) that there is a significant difference between female and male in relation to the weights of rats. Similarly, under this same level, there is a significant difference between treatments [control (0) vs. high (1)] and [control (0) vs. low (2)].

Table 10 gives the AIC, BIC and GD statistics for some regressions. The results indicate that the OLLExGa regression has the smallest values of these statistics among all fitted regressions. So, it could be chosen as the more suitable model to these data. The LR statistic for testing $H_0 : \tau = 1$ versus $H_1 : \tau \neq 1$, i.e. to compare the OLLExGa and ExGa regressions, is $w = 9.6831$ ($p-$value $= 0.0019$). This $p$-value indicates that the first regression yields the best fit to the weights of rat pups data.

We use the **R** software to compute $LD_i(\boldsymbol{\theta})$ and $GD_i(\boldsymbol{\theta})$ in the diagnostic analysis discussed in Section 4. The results of such influence measures index plots are displayed in Figure 9. The plots reveal that the cases ♯66, ♯298, ♯300 and ♯305 are possible influential observation.

In addition, Figure 10(a) gives the index plot of the qrs for the fitted model. The observation ♯66 is just one out of the range $[-3, 3]$. Hence, there is no evidence against the model assumptions. We present the normal plot for the qrs with a generated envelope in Figure 10(b) to detect possible departures from these assumptions and outliers. This plot shows that the fitted OLLExGa regression provides a good fit to the current data, since only two points are outside the envelope.

In order to assess whether the model fits the data appropriately, the empirical cdf and estimated cdf of the OLLExGa regression are plotted in Figure 10(c) for the sex explanatory variable. We can note a significant difference between female and male individuals in relation to the weights of the rats.

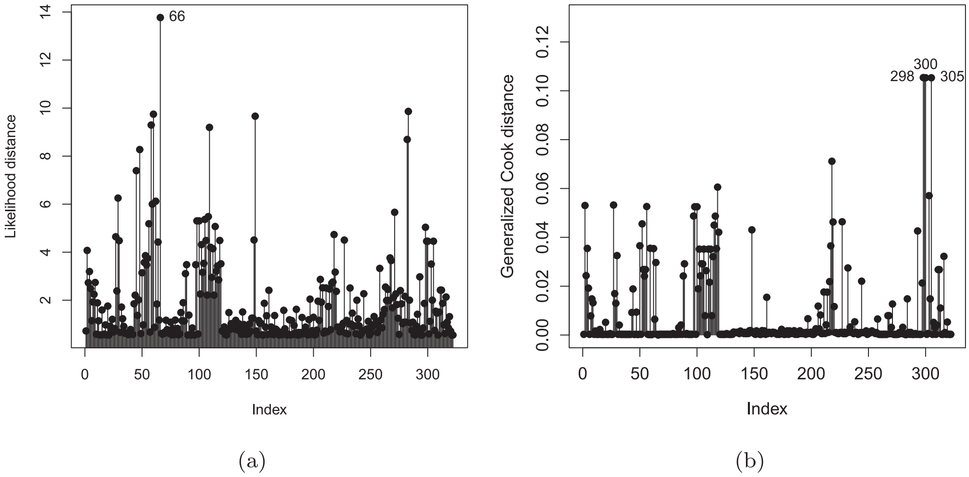


**Figure 9.** Index plot for $\boldsymbol{\theta}$: (a) $LD_i(\boldsymbol{\theta})$ (likelihood distance) and (b) $GD_i(\boldsymbol{\theta})$ (generalized Cook's distance).

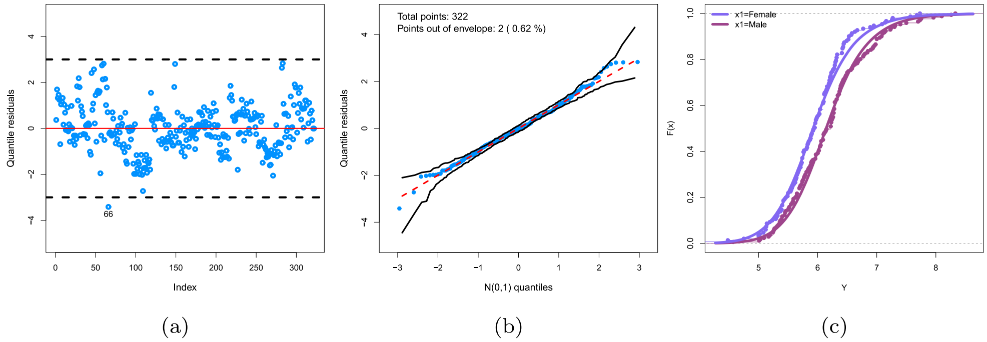


**Figure 10.** (a) Index plot of the $qr_i$. (b) Normal probability plot with envelope for the $qr_i$. (c) Estimated cdf from the fitted OLLExGa regression to the weights of rat pups data and the empirical cdf.

### *5.3. Application 3: degrees brix of yacon data*

The data from the third application (yacon data) were taken from the package *agricolae* [4] available in R software having as source: CIP, Experimental field, 2003. The data were kindly provided by Ivan Manrique and Carolina Tasso. The third application considers the OLLExGa regression with two systematic components for $\mu$ and $\sigma$. The data ($n = 432$) refer to a native plant of the Peruvian Andes called *yacon* (*Smallanthus sonchifolius*), which is a common plant in the country. We use the covariable location (Cajamarca, Lima, Oxapampa) to verify how much this variable explains the response variable *degrees brix* (a numerical scale that measures the density or sugar concentration of solutions). The data belong to the International Potato Center in Lima (Peru).

The variables for the regression analysis are ($i = 1, \ldots, 432$):

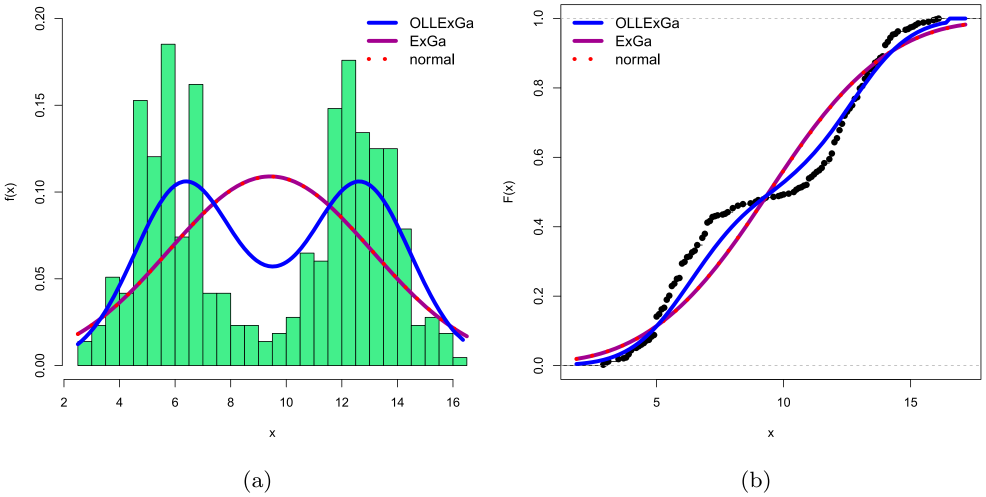
- $y_i$: degrees brix of *yacon* (response variable);
- $x_{i1}$: locale ($0 =$ Cajamarca, $1 =$ Lima, $2 =$ Oxapampa) (two dummy variables).

**Table 11.** Descriptive statistics for degrees brix of *yacon*.

| Mean | Median | SD | Skewness | Kurtosis | Min. | Max. |
|---|---|---|---|---|---|---|
| 9.4310 | 10.4500 | 3.6674 | −0.0492 | −1.5410 | 2.9000 | 16.1000 |

**Table 12.** MLEs, SEs and AIC, BIC and GD values for the fitted models to degrees brix data.

| Model | $\mu$ | $\log(\sigma)$ | $\log(\nu)$ | $\tau$ | AIC | BIC | GD |
|---|---|---|---|---|---|---|---|
| OLLExGa | 9.5262 | −0.1805 | −3.892 | 0.1194 | 2149.156 | 2165.652 | 2141.379 |
| | (0.0448) | (0.0079) | (4811.252) | (0.0065) | | | |
| ExGa | 9.1299 | 1.2943 | −1.212 | 1 | 2353.742 | 2365.950 | 2347.744 |
| | (0.9904) | (0.0400) | (3.276) | (-) | | | |
| normal | 9.4313 | 1.2983 | (-) | (-) | 2351.722 | 2359.859 | 2347.722 |
| | (0.1762) | (0.0340) | (-) | (-) | | | |



(a) (b)

**Figure 11.** (a) Estimated densities of the OLLExGa, ExGa and normal models for power generation data. (b) Estimated cumulative functions of the OLLExGa, ExGa and normal models and the empirical cdf for degrees brix data.

In the first part, we perform a univariate analysis considering only the response variable. Table 11 gives the descriptive analysis of the response variable, where the mean and median are 9.431 and 10.450, respectively. Then, the data have an asymmetric shape.

We provide in Table 12 the AIC, BIC and GD measures for three fitted regressions which indicate that the OLLExGa regression has the lowest values of these statistics. So, it could be chosen as the best regression for these data.

The LR statistic for testing the hypotheses $H_0 : \tau = 1$ versus $H_1 : \tau \neq 1$, i.e. to compare the OLLExGa and ExGa regressions, is $w = 206.3659$ ($p$—value $=< 0.0001$). Clearly, the proposed distribution outperforms the ExGa distribution based on the value of this statistic. We display in Figure 11(a), the histogram of the data and the plots of the fitted OLLExGa, ExGa and normal densities. We given in Figure 11(b) the plots of the empirical cdf and fitted OLLExGa, ExGa and normal cumulative distributions. The plots confirm that the OLLExGa distribution provides a better fit to these data. The plot in Figure 11(a) reveals that the degrees brix histogram has a bimodality shape, where the ExGa and normal distributions cannot have this shape.

*Regression analysis with two systematic components*

We consider the OLLExGa regression with two systematic components

$$\mu_i = \beta_{10} + \beta_{11} x_{i1(0-1)} + \beta_{12} x_{i1(0-2)}$$

and

$$\sigma_i = \exp(\beta_{20} + \beta_{21} x_{i1(0-1)} + \beta_{22} x_{i1(0-2)}), \quad i = 1, \ldots, 432.$$

Table 13 gives the MLEs, SEs and their $p$-values for this fitted regression. The figures in this table indicate that all covariables are significant in the two systematic components for a 5% significance level. Thus, there is a significant difference between the localities [Cajamarca (0) vs Lima (1)] and [Cajamarca (0) versus Oxapampa (2)] in relation to the degrees brix of *yacon*.

Further, the figures in Table 14 indicate that the OLLExGa regression has the lowest AIC, BIC and GD values among those of the fitted regressions. So, it could be chosen as the best model. We compare the OLLExGa and ExGa regressions using LR statistic. This statistic for testing the hypotheses $H_0 : \tau = 1$ versus $H_1 : H_0$ is not true, is $w = 26.7174$ ($p-$value $=<$ 0.001), which yields favorable indications toward to the OLLExGa regression (see Table 15).

We use the **R** software to compute the $LD_i(\boldsymbol{\theta})$ and $GD_i(\boldsymbol{\theta})$ measures in the diagnostic analysis presented in Section 4. The results of such influence measures plots are displayed in Figure 12. These plots reveal that the cases ♯151, ♯173, ♯205, ♯297, ♯303, ♯323, ♯378 and ♯408 are possible influential observations.

**Table 13.** MLEs, SEs and $p$-values for the OLLExGa regression fitted to the degrees brix of *yacon* data.

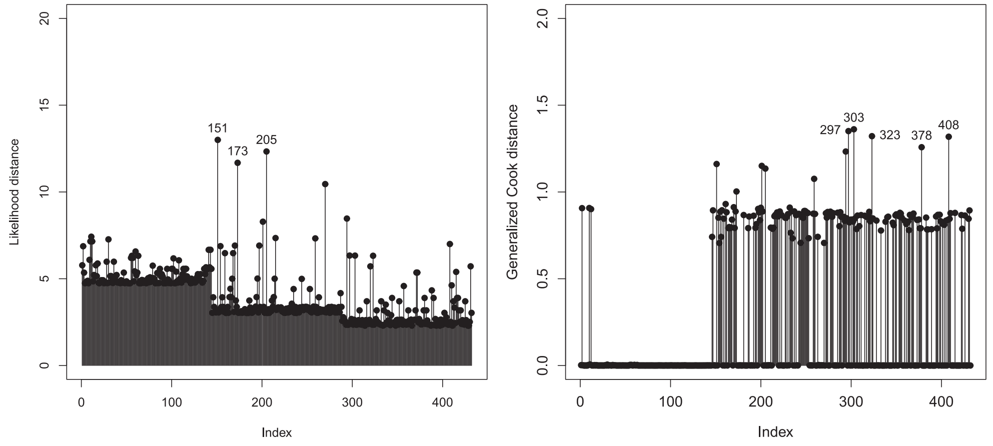
| Parameter | Estimate | SE | $p$-Value |
|---|---|---|---|
| $\beta_{10}$ | 9.7882 | 0.2093 | $< 0.001$ |
| $\beta_{11}$ | −3.8355 | 0.1996 | 0.0021 |
| $\beta_{12}$ | 2.7567 | 0.2251 | $< 0.001$ |
| $\beta_{20}$ | 0.3485 | 0.1530 | 0.0233 |
| $\beta_{21}$ | −0.8574 | 0.0639 | 0.0021 |
| $\beta_{22}$ | −1.2253 | 0.0623 | $< 0.001$ |
| $\log(\nu)$ | −3.5055 | 0.1513 | |
| $\tau$ | 0.2890 | 0.0659 | |

**Table 14.** Goodness-of-fit measures for degrees brix of *yacon* data.

| Model | AIC | BIC | GD |
|---|---|---|---|
| OLLExGa | 1714.103 | 1746.651 | 1698.103 |
| normal | 1736.773 | 1761.184 | 1724.773 |
| ExGa | 1738.821 | 1767.300 | 1724.821 |

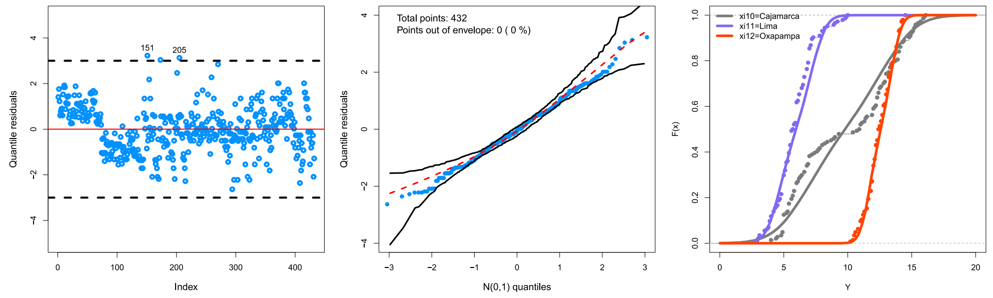
**Table 15.** LR tests for degrees brix of *yacon* data.

| Models | Hypotheses | Statistic $w$ | $p$-value |
|---|---|---|---|
| OLLExGa vs ExGa | $H_0 : \tau = 1$ vs $H_1 : H_0$ is false | 26.7174 | $< 0.001$ |



(a) (b)

**Figure 12.** Index plot for $\boldsymbol{\theta}$: (a) $LD_i(\boldsymbol{\theta})$ (likelihood distance) and (b) $GD_i(\boldsymbol{\theta})$ (generalized Cook's distance).



(a) (b) (c)

**Figure 13.** (a) Index plot of the $qr_i$. (b) Normal probability plot with envelope for $qr_i$. (c) Estimated cdf from the fitted OLLExGa regression model and empirical cdf (for the localities) for the degrees brix of *yacon* data.

In addition, Figure 13(a) gives plots of the qrs for the fitted regression which indicate a random behavior of these residuals and that the cases ♯151 and ♯205 are out of the range $[-3, 3]$. The normal plot for the qrs with a generated envelope is displayed in Figure 13(b). These plots confirm that the model assumptions hold and that the fitted regression explain the data of the degrees brix of the *yacon* plant to certain localities of Peru.

A graphical comparison among the localities (Cajamarca, Lima, Oxapampa) is give in Figure 13(c). These plots provide the empirical cdf and the estimated cdf of the OLLExGa regression model. It is clear from these plots that the OLLExGa regression presents a good fit. We can also note that there is a significant difference among localities in relation to the degrees brix of *yacon*. In summary, the OLLExGa regression outperforms the ExGa and normal regressions irrespective of the criteria and then it can be effectively adopted to fit these data.

## 6. Concluding remarks

We propose a new four parameter *odd log-logistic exponential Gaussian* (OLLExGa) distribution which allows modeling of bimodal data without requiring mixtures of distributions. We also define a more realistic regression based on the new distribution for modeling agriculture real data. It is an important extension of some well-known regression models. Some Monte Carlo simulation studies investigate the accuracy of the normal approximation for the maximum-likelihood estimators of the model parameters. Diagnostic measures and quantile residuals are investigated to verify the sensitivity of these estimators. We prove empirically the usefulness of the proposed models by means of three real data sets applied in agricultural experiments.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil.

## ORCID

*Gauss Moutinho Cordeiro* http://orcid.org/0000-0002-3052-6551
*Edwin Moises Marcos Ortega* http://orcid.org/0000-0003-3999-7402

## References

[1] A.C. Atkinson, *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Clarendon Press, Oxford, 1985.

[2] R.D. Cook, *Assessment of local influence*, J. R. Stat. Soc. Ser. B Methodol. 48 (1986), pp. 133–155.

[3] R.D. Cook and S. Weisberg, *Residuals and Influence in Regression*, Chapman and Hall, New York, 1982.

[4] F. de Mendiburu and M.F. de Mendiburu, Package 'agricolae'. R Package, 2019, pp. 1–2.

[5] P. Dunn and G. Smyth, *Randomized quantile residuals*, J. Comput. Graph. Stat. 5 (1996), pp. 236–244.

[6] J.U. Gleaton and J.D. Lynch, *Properties of generalized log-logistic families of lifetime distributions*, J. Probab. Statist. Sci. 4 (2006), pp. 51–64.

[7] A. Golubev, *Exponentially modified Gaussian (EMG) relevance to distributions related to cell proliferation and differentiation*, J. Theor. Biol. 262 (2010), pp. 257–266.

[8] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals Series, and Products*, Academic Press, San Diego, CA, 2000.

[9] G.G. Hamedani, E. Altun, M.Ç. Korkmaz, H.M. Yousof, and N.S. Butt, *A new extended G family of continuous distributions with mathematical properties, characterizations and regression modeling*, Pakistan J. Statist. Oper. Res. 14 (2018), pp. 737–758.

[10] E.M. Hashimoto, E.M.M. Ortega, G.M. Cordeiro, V.G. Canchoand, and C. Klauberg, *Zero-spiked regression models generated by gamma random variables with application in the resin oil production*, J. Stat. Comput. Simul. 89 (2018), pp. 52–70.

[11] Y. Kalambet, Y. Kozmin, k. Mikhailova, I. Nagaev, and P. Tikhonov, *Reconstruction of chromatographic peaks using the exponentially modified Gaussian function*, J. Chemom. 25 (2011), pp. 352–356.

[12] M.Ç. Korkmaz, E. Altun, M. Alizadeh, and H.M. Yousof, *A new flexible lifetime model with log-location regression modeling, properties and applications*, J. Stat. Manag. Syst. 22 (2019), pp. 871–891.

[13] M.Ç. Korkmaz, E. Altun, H.M. Yousof, and G.G. Hamedani, *The odd power Lindley generator of probability distributions: Properties, characterizations and regression modeling*, Int. J. Statist. Probab. 8 (2019), pp. 70–89.

[14] M.Ç. Korkmaz, G.M. Cordeiro, H.M. Yousof, R.R. Pescim, A.Z. Afify, and S. Nadarajah, *The Weibull Marshall-Olkin family: Regression model and application to censored data*, Comm. Stat. Theory Methods 48 (2019), pp. 4171–4194.

[15] S.V. Magnitskiy, C.C. Pasian, M.A. Bennett, and J.D. Metzger, *Effects of soaking cucumber and tomato seeds in paclobutrazol solutions on fruit weight, fruit size, and paclobutrazol level in fruits*, HortScience 41 (2006), pp. 1446–1448.

[16] J.D. Maguire, *Speed of germination aid in selection and evaluation for seedling emergence and vigor*, Crop Sci. 2 (1962), pp. 176–177.

[17] H.T.B Oliveira, E.C. Pereira, V. Mendonça, R.M.S. Silva, G.A. Leite, and L.L.G.R. Dantas, *Produção e qualidade de frutos de mangueira tommy aktins sob doses de paclobutrazol*, Agropecuária Científica no Semiárido 10 (2015), pp. 89–92.

[18] E.M. Palmer, T.S. Horowitz, A. Torralba, and J.M. Wolfe, *What are the shapes of response time distributions in visual search*, J. Exp. Psych. Hum. Percep. Perform. 37 (2011), pp. 58–71.

[19] M. Panobianco, R.D. Vieira, F.C. Krzyzanowski, and J.F. Neto, *Electrical conductivity of soybean seed and correlation with seed coat lignin content*, Seed Sci. Technol. 27(3) (1999), pp. 945–949.

[20] J. C. Pinheiro and D. M. Bates, *Mixed-Effects Models in S and S-PLUS*, Springer, New York, NY, 2006.

[21] F. Prataviera, E.M.M. Ortega, G.M. Cordeiro, and A.S. Braga, *The heteroscedastic odd log-logistic generalized gamma regression model for censored data*, Comm. Statist. – Simulation Comput. 48 (2018a), pp. 1–25.

[22] F. Prataviera, E.M.M. Ortega, G.M. Cordeiro, R.R. Pescim, and B.A.W Verssani, *A new generalized odd log-logistic flexible Weibull regression model with applications in repairable systems*, Reliab. Eng. Syst. Safety. 176 (2018b), pp. 13–26.

[23] F. Prataviera, J.C.S. Vasconcelos, G.M. Cordeiro, E.M. Hashimoto, and E.M.M. Ortega., *The exponentiated power exponential regression model with different regression structures: application in nursing data*, J. Appl. Statist. 46 (2019), pp. 1792–1821.

[24] É.M.D. Rezende, J.A. Oliveira, E.R. Carvalho, A.D.C.S. Clemente, and G.E. Oliveira, *Physiological quality of tomato seeds treated with polymers in combination with paclobutrazol*, J. Seed Sci. 39 (2017), pp. 338–343.

[25] A. Seleguini, M.J.A. Faria Júnior, K.S.S. Bennet, O.L. Lemos, and S. Seno, *Estratégias para produção de mudas de tomateiro utilizando paclobutrazol*, Semina: Ciências Agrárias 34 (2013), pp. 539–548.

[26] J.C. Souza Vasconcelos, G.M. Cordeiro, E.M. Ortega, and E.G. Araújo, *The new odd log-Logistic generalized inverse Gaussian regression model*, J. Probab. Statist. 2019 (2019), pp. 1–13.

[27] D.M. Stasinopoulos and R.A. Rigby, *Generalized additive models for location scale and shape (GAMLSS) in R*, J. Statist. Softw. 23 (2007), pp. 1–46.

[28] L. Taiz and E. Zeiger, *Fisiologia Vegetal*, 5th ed., Editora Artmed, Semina: Porto Alegre, 2013.

[29] D.R. Tyson, S.P. Garbett, P.L. Frick, and V. Quaranta, *Fractional proliferation: A method to deconvolve cell population dynamics from single-cell data*, Nat. Methods 9 (2012), pp. 923–928.

[30] F.C. Xie and B.C. Wei, *Diagnostics analysis in censored generalized poisson regression model*, J. Stat. Comput. Simul. 77 (2007), pp. 695–708.

## Appendix

In this Appendix, we derive useful expansions for $F(y)$ in (3) and $f(y)$ in (4) to find structural properties for the proposed distribution. First, we consider the following power series for $\eta(y)^{\tau}$ for any real $\tau$ since $\eta(y) \in (0,1)$

$$\eta(y)^{\tau} = \sum_{k=0}^{\infty} c_k \eta(y)^k, \tag{A1}$$

where

$$c_k = c_k(\tau) = \sum_{j=k}^{\infty} (-1)^{k+j} \binom{\alpha}{j} \binom{j}{k}.$$

For any real $\tau$, we use the generalized binomial expansion

$$\left[1 - \eta(y)\right]^{\tau} = \sum_{k=0}^{\infty} (-1)^k \binom{\alpha}{k} \eta(y)^k. \tag{A2}$$

Inserting (A1) and (A2) in Equation (3), we have

$$F(y) = \frac{\sum_{k=0}^{\infty} c_k \eta(y)^k}{\sum_{k=0}^{\infty} d_k \eta(y)^k},$$

where $d_k = d_k(\tau) = c_k(\tau) + (-1)^k \binom{\tau}{k}$ (for $k \geq 0$). The ratio of these two power series is expressed as

$$F(y) = \sum_{k=0}^{\infty} w_k \eta(y)^k, \tag{A3}$$

where the coefficients $w_k$'s (for $k \geq 0$) are determined from the recurrence equation

$$w_k = w_k(\tau) = d_0^{-1} \left( c_k - \sum_{r=1}^{k} d_r \, w_{k-r} \right).$$

In practical terms, we need only six terms in (A3) to achieve good approximations for $F(y)$. By differentiating (A3), we obtain he density of $Y$

$$f(y) = \sum_{k=0}^{\infty} w_{k+1} \pi_{k+1}(y), \tag{A4}$$

where $\pi_{k+1}(y) = (k+1)\eta(y)^k g_{\mu,\sigma,\nu}(y)$ is the *exponentiated exponential Gaussian* (EEGa) density function with power parameter $k+1$ (for $k \geq 0$).

Further, we can write the cdf of the ExGa distribution from (1) as

$$\eta(y) = K \left[ \frac{1}{2} + \int_{-p}^{v} \mathrm{e}^{-\mathrm{q}\mathrm{z}} \mathrm{erf}\left( \frac{\mathrm{z}}{\sqrt{2}} \right) \mathrm{d}\mathrm{z} \right],$$

where $K = \nu^{-1}\,\mathrm{e}^{-\sigma^2/(2\nu^2)}$, $q = \sigma/\nu$, $p = (\mu\nu + \sigma^2)/(\nu\sigma)$, $v = v(y) = (y-\mu)/\sigma - \sigma/\nu$ and $\mathrm{erf}(z) = 2\Phi(\sqrt{2}z) - 1$ is the error function.

By expanding the last integral (say $I$) in Taylor series using Mathematica at $v = 0$,

$$I(v) = \sum_{i=0}^{\infty} a_i^{\star} v^i,$$

where

$$a_0^{\star} = q^{-1} \left[ \mathrm{e}^{-\mathrm{q}^2/2} \mathrm{erf}\left( \frac{\mathrm{q}-\mathrm{p}}{\sqrt{2}} \right) + \mathrm{e}^{\mathrm{q}^2/2} \mathrm{erf}\left( \frac{\mathrm{q}}{\sqrt{2}} \right) - \mathrm{e}^{-\mathrm{q}\mathrm{p}} \mathrm{erf}\left( \frac{\mathrm{p}}{\sqrt{2}} \right) \right],$$

$a_1^{\star} = 0$, $a_2^{\star} = 1/\sqrt{2\pi}$, $a_3^{\star} = (-q/3)\sqrt{2/\pi}$ and $a_4^{\star} = (3q^2-1)/(12\sqrt{2\pi})$, etc. Then, we can rewrite $\eta(y)$ as

$$\eta(y) = \sum_{i=0}^{\infty} a_i v^i,$$

where $a_0 = a_0^\star + K/2$ and $a_i = Ka_i^\star$ for $i \geq 1$. We can take at most four to six terms in this power series to approximate adequately $\eta(y)$.

A power series raised to a positive integer power $k$ is given by [8], Section 0.314

$$\eta(y)^k = \left(\sum_{i=0}^{\infty} a_i v^i\right)^k = \sum_{i=0}^{\infty} c_{k,i} v^i, \tag{A5}$$

where the coefficients $c_{k,i}$ (for $i = 1, 2, \ldots$) are determined recursively from the equation

$$c_{k,i} = (ia_0)^{-1} \sum_{m=1}^{i} [m(k+1) - i] a_m c_{k,i-m},$$

and $c_{k,0} = a_0^k$. The coefficient $c_{k,i}$ can be given explicitly in terms of the coefficients $a_i$'s, although it is not necessary for programming numerically our expansions in any algebraic or numerical software.

Hence, the density function of $Y$ can be expressed from (A4) and (A5) as

$$f(y) = \sum_{i=0}^{\infty} s_i \left[(y-\mu)/\sigma - \sigma/\nu\right]^i g_{\mu,\sigma,\nu}(y), \tag{A6}$$

where $s_i = w_1 + \sum_{k=1}^{\infty} (k+1) w_{k+1} c_{k,i}$. In applications, the index $k$ runs at most up to five.

We can obtain some mathematical properties of the OLLExGa distribution using (A6) from those of the ExGa distribution with small values for $i$ such as four. For example, the $n$th ordinary moment of $Y$ can be expressed as a linear combination of those ordinary moments of $W$ with orders $n$, $n+1$ up to $n+4$.

Let $T(\cdot)$ be any integrable function on a real line and $Q_W(u)$ be the qf of the ExGa distribution. We can write

$$\int_{-\infty}^{\infty} T(y) f(y)\, \mathrm{d}y = \sum_{i=0}^{\infty} s_i \int_0^1 T[Q_W(u)] \left\{[Q_W(u) - \mu]/\sigma - \sigma/\nu\right\}^i \mathrm{d}u. \tag{A7}$$

Hence, based on (A7), several mathematical quantities of the OLLExGa distribution can be computed numerically from linear combinations of integrals over $(0,1)$ of adequate functions involving only the qf of the ExGa distribution.