mathematics

MDPI

Article

Modelling Asymmetric Data by Using the Log-Gamma-Normal Regression Model

Roger Tovar-Falón [1,*,†], Guillermo Martínez-Flórez [1,†] and Heleno Bolfarine [2,†]

[1] Departamento de Matemáticas y Estadística, Facultad de Ciencias Básicas, Universidad de Córdoba, Monteria 230002, Colombia; guillermomartinez@correo.unicordoba.edu.co
[2] Departamento de Estatística, Universidade de São Paulo, Sao Paulo 05508-090, Brazil; hbolfar@ime.usp.br
* Correspondence: rjtovar@correo.unicordoba.edu.co
† These authors contributed equally to this work.

Abstract: In this paper, we propose a linear regression model in which the error term follows a log-gamma-normal (LGN) distribution. The assumption of LGN distribution gives flexibility to accommodate skew forms to the left and to the right. Kurtosis greater or smaller than the normal model can also be accommodated. The regression model for censored asymmetric data is also considered (censored LGN model). Parameter estimation is implemented using the maximum likelihood approach and a small simulation study is conducted to evaluate parameter recovery. The main conclusion is that the approach is very much satisfactory for moderate and large sample sizes. Results for two applications of the proposed model to real datasets are provided for illustrative purposes.

Keywords: log-gamma-normal distribution; linear regression models; asymmetric data; censored data; maximum likelihood estimators

MSC: 60E05

Citation: Tovar-Falón, R.; Martínez-Flórez, G.; Bolfarine, H. Modelling Asymmetric Data by Using the Log-Gamma-Normal Regression Model. *Mathematics* **2022**, *10*, 1199. https://doi.org/10.3390/math10071199

Academic Editor: Ion Mihai

Received: 8 March 2022
Accepted: 1 April 2022
Published: 6 April 2022

1. Introduction

Regression models are one of the main statistical techniques frequently used in data analysis in any area of knowledge, especially when there is interest in studying the relationship between a dependent variable (response) and two or more independent (explanatory) variables. In this sense, a regression model with a response variable following a normal distribution is perhaps best known in the literature, and could be considered one of the most widely used; however, the assumption of normality may not be adequate in the dataset under analysis, since these may present degrees of skewness or kurtosis that are not within the range covered by the normal model. Consequently, inferences made from the fitted model may not have statistical validity, and erroneous conclusions may be reached. A solution to the problem of the assumption of normality of the variable of interest is the use of transformations, although it is well known that this solution makes it difficult to interpret results since data are not in the original measurement scale. As an alternative to this issue, many authors have introduced new family distributions that are capable of capturing degrees of skewness and kurtosis greater than those that the normal distribution can capture.

One of the most important works in the context of data with a high degree of asymmetry is Azzalini [1], which is known in the statistical literature as the skew-normal (SN) model. The main characteristic of the SN model is its ability to fit degrees of asymmetry (on the left and right) greater than those of the normal model; however, it is not the best model in terms of capturing high degrees of kurtosis. Relating to the latter, the power-normal (PN) model introduced by Durrans [2] has the particularity of fitting data with a higher degree of kurtosis than the normal and SN model but with less range of asymmetry. The SN and PN models have been studied extensively by many authors, and different

extensions of this model have been considered. In Gupta and Gupta [3] for example, the authors showed the existing practical problems when the asymmetry parameter of the SN model is estimated, and they proposed an alternative model named the PN model. The authors also investigated the closeness between the proposed model and the SN model. In Pewsey et al. [4], the authors presented the general results of the likelihood-based inference for the family of power distributions, with particular emphasis on the case of the PN model, complementing the work of Gupta and Gupta [3]. In Martínez-Flórez et al. [5], the authors introduced a new model that generalized both the SN model by Azzalini [1] and the PN model by Durrans [2]. The new model, which is called power-skew-normal (PSN), has the particularity of fitting data with degrees of asymmetry greater than those of the SN model, and is also capable of capturing degrees of kurtosis greater than those of the PN model. Furthermore, the authors showed that the information matrix of the new model is non-singular, which permits carrying out hypothesis tests on the asymmetry parameters based on likelihood-ratio statistics. On the other hand, Martínez-Flórez et al. [6] generalized the log-normal (LN) model from the SN and PN models. In addition, these new proposals contain the LN model as a particular case, and they are more flexible regarding skewness and kurtosis to fit positive data.

Alternatives for fitting asymmetric data with a high degree of kurtosis were reported by other authors such as Tovar-Falón et al. [7], who introduced a new model that generalizes the skew-t model of Azzalini and Capitanio [8] and power-t of Zhao and Kim [9]. Here, the inference was carried out from a classical perspective using the maximum likelihood method. This new model also has, as particular cases, the PSN, SN, PN, Student-t and normal models. In Tung et al. [10], the authors considered a mixture class of log-F distributions to characterize asymmetric distributions by integrating it into a pH acceleration model. The authors studied the impact of the new model in the presence of misspecification of particle size distribution.

Models for asymmetric data with high degrees of skewness and kurtosis, and presenting more than one mode and censored data, were also considered. More details about these topics can be found in Martínez-Flórez et al. [11] and Martínez-Flórez et al. [12], respectively. In addition, all the aforementioned models are easily extensible to the situations of regression models, including cases in which the data show censoring in some value; see Sahu et al. [13], Martínez-Flórez et al. [14].

In Amini et al. [15], the authors introduced a new family of distributions useful for modelling asymmetric data. This new family of continuous distributions is generated by a distribution F and two positive real parameters δ and γ, which control the skewness and tail weight of the distribution. The probability density function (PDF) of this family is given by

$$g(x;\delta,\gamma) = \frac{\gamma^{\delta}}{\Gamma(\delta)}[-\ln F(x)]^{\delta-1}[F(x)]^{\gamma-1}f(x), \tag{1}$$

where $\delta,\gamma \in \mathbb{R}^{+}$ and $\Gamma(\cdot)$ is the complete gamma function, $F(\cdot)$ is the cumulative distribution function (CDF) of X, and $f(\cdot)$ is the associated PDF. In this work, the authors studied the main properties of the distribution and addressed the estimation process of the unknown parameters of the model using the likelihood approach.

From the $F(x)$ generator, the authors studied some particular properties of the family, among which are the exponential, Weibull, power, Pareto, extreme value and Gumbel distributions. If $\delta = 1$ and $F(\cdot) = \Phi(\cdot)$ and $f(\cdot) = \phi(\cdot)$ in model (1), i.e., the CDF and PDF of the standard normal distribution, respectively, the model in (1) is reduced to the PN model by Durrans [2]. Hence, the model in (1) is an extension of the PN model. In Cordeiro et al. [16], the authors studied in detail the properties of the log-gamma-generated family of distributions introduced by Amini et al. [15] and presented some applications of this family. Other particular cases of the model introduced by Amini et al. [15] correspond to the generalized gamma and log-gamma distributions, which have been extensively studied by many authors; see Prentice [17], Lawless [18], Young and Bakir [19], Ortega et al. [20,21] among others.

The main goal of this article is to focus on the study of the regression model under the assumption that the errors follow a log-gamma-normal (LGN) distribution, which is obtained by taking $F(x) = \Phi(x)$ and $f(x) = \phi(x)$ in model (1). We also consider the case of the regression model for censored data, and we conduce the parameter estimation using the maximum likelihood approach and its large sample properties.

Although there are many works in the literature related to the generalized log-gamma distribution, our proposal is based on the family of distributions presented by Amini et al. [15], which is known in the literature as log-gamma-generated. For the case of this family, we focus on the case in which the generating function is the normal distribution and the distribution called log-gamma-normal is obtained, and this distribution does not correspond to the distributions previously mentioned. In our proposal, we change the assumption that the errors in the multiple linear regression model follow a normal distribution to that of errors with a log-gamma-normal distribution. It is also important to note that the generalized log-gamma and log-gamma distributions are also particular cases of the family introduced by Amini et al. [15] (assuming a gamma distribution instead of the standard normal distribution), but in our proposal we do not consider these cases.

In addition to carrying out the estimation process of the parameters in the model, we present two applications using real datasets. The first dataset was previously analyzed by Zhang and Davidian [22], and the second dataset is related to a study on the abundance of beryllium scaled to the Sun's abundance. For the particular case of these datasets, the model fits well, and therefore we can conclude that, apart from the existence of statistical literature for the analysis of asymmetric data, our proposal is a viable alternative that competes with existing models. The main contribution of this model is that the trend of the dataset under examination is better explained using a model with log-gamma type errors instead of one with asymmetric errors using another distribution.

The article is organized as follows. In Section 2, we define the family of LGN distributions and discuss some of its properties. In Section 3, the LGN regression model is defined, and its properties studied. The inference is implemented using the maximum likelihood approach. The censored LGN model for dealing with censored data by maximum likelihood estimation is discussed in Section 4. The results of a small-scale simulation study reveal the good performance of the estimation approach in Section 5. In Section 6, two real data applications are considered, revealing that the datasets in question are better fitted by LGN model than PN and models.

2. Log-Gamma-Normal Distribution

In this section, we define the LGN model, which is obtained from the family given in (1) by taking the CDF of the standard normal distribution, and we study some basic properties.

Definition 1. *The random variable X is said to have a LGN distribution, if X has PDF given by*

$$f_{LGN}(x;\delta,\gamma) = \frac{\gamma^{\delta}}{\Gamma(\delta)}[-\ln\Phi(x)]^{\delta-1}[\Phi(x)]^{\gamma-1}\phi(x), \quad x \in \mathbb{R}, \tag{2}$$

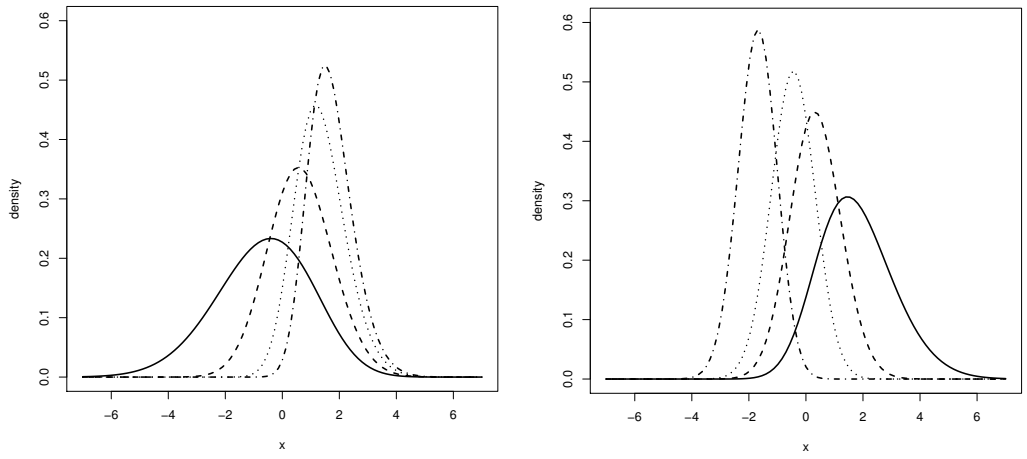
where $\delta, \gamma \in \mathbb{R}^+$, $\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}du$ *is the gamma function, and the functions* $\phi(\cdot)$ *and* $\Phi(\cdot)$ *are the PDF and CDF of the standard normal distribution, respectively.*

A random variable with LGN distribution is shortly denoted by $X \sim \text{LGN}(\delta,\gamma)$. One can note that the function (2) is a proper PDF since $f_{LGN}(x;\delta,\gamma) \geq 0$ for all $x \in \mathbb{R}$ and $\delta, \gamma \in \mathbb{R}^+$. Thus, letting $y = -\ln\Phi(x)$, it follows that

$$\int_{\mathbb{R}} \frac{\gamma^{\delta}}{\Gamma(\delta)}[-\ln\Phi(x)]^{\delta-1}[\Phi(x)]^{\gamma-1}\phi(x)dx = \int_0^{+\infty} \frac{\gamma^{\delta}}{\Gamma(\delta)}y^{\delta-1}e^{-\gamma y}dy = 1.$$

Figure 1 depicts some shapes of LGN distribution for some selected values of the parameters δ and γ. It can be seen that the parameters δ and γ affect both the skewness

and kurtosis of the model, and hence, the LGN distribution is more flexible for fitting data that may be skewed as well as having thinner or thicker tails than the normal, SN and PN distributions.



(**a**) (**b**)

Figure 1. PDF of the LGN distribution: (**a**) $\delta = 0.6$ and $\gamma = 0.30$ (dotted line), 0.60 (dashed line), 1.0 (dotted–dashed line), 3.0 (long dashed line) and 6.0 (solid line). (**b**) $\gamma = 1.5$ and $\delta = 0.30$ (dotted line), 0.60 (dashed line), 1.0 (dotted–dashed line), 2.0 (long dashed line) and 5.0 (solid line).

The LGN distribution reduces to some specific distributions as special cases for specified values of the parameters δ and γ; some of them are available in the literature and have been widely studied.

Proposition 1. *Let* $X \sim LGN(\delta, \gamma)$
(i) *if* $\delta = \gamma = 1$, *then* $X \sim N(0, 1)$,
(ii) *if* $\delta = 1$, *then* $X \sim PN(\gamma)$,
(iii) *if* $\delta = 1$ *and* $\gamma = 2$, *then* $X \sim SN(1)$.

Proof. Demonstration of (i)–(iii) is immediate from the definition of LGN distribution. □

2.1. Moments

Measures of skewness and kurtosis can be given from the moments of the LGN distribution. The following proposition gives an expression of the rth moment of the random variable $X \sim \text{LGN}(\delta, \gamma)$ which does not have a closed form.

Proposition 2. *Let* $X \sim LGN(\delta, \gamma)$ *then*

$$\mathbb{E}\left[X^k\right] = \mathbb{E}\left[\left(\Phi^{-1}(e^{-W})\right)^k\right], \quad \textit{for} \quad k = 1, \ldots, n, \tag{3}$$

where $\Phi^{-1}(\cdot)$ *is the inverse of the CDF* $\Phi(\cdot)$ *and the random variable W follows a gamma distribution with parameters* δ *and* γ.

Proof. We have by definition that

$$\mathbb{E}\left[X^k\right] = \int_{\mathbb{R}} x^k \frac{\gamma^\delta}{\Gamma(\delta)} [-\ln \Phi(x)]^{\delta-1} [\Phi(x)]^{\gamma-1} \phi(x) dx.$$

Letting $W = -\ln \Phi(X)$, then $X = \Phi^{-1}(e^{-W})$, it follows that

$$\mathbb{E}\left[X^k\right] = \int_0^{+\infty} \left(\Phi^{-1}(e^{-w})\right)^k \frac{\gamma^\delta}{\Gamma(\delta)} w^{\delta-1} e^{-\gamma w} dw,$$

which is the expected value of the function $\left[\Phi^{-1}(e^{-W})\right]^k$, where W follows a Gamma(δ, γ) distribution. □

Based on moments (3), one can obtain the skewness $(\sqrt{\beta_1})$ and the kurtosis (β_2) coefficients of the LGN model using the following expressions

$$\sqrt{\beta_1} = \frac{\mu_3 - 3\mu_1\mu_2 + 2\mu_1^3}{(\mu_2 - \mu_1^2)^{3/2}},$$

and

$$\beta_2 = \frac{\mu_4 - 4\mu_1\mu_3 + 6\mu_2\mu_1^2 - 3\mu_1^4}{(\mu_2 - \mu_1^2)^2}$$

respectively, where $\mu_k = \mathbb{E}\left[X^k\right]$ for $k = 1, \ldots, 4$. The skewness and kurtosis coefficients for values of δ and γ ranging between 0.1 and 200 were calculated using numerical integration with an integrate function of R Development Core Team [23] for LGN model. It was found that $\sqrt{\beta_1} \in [-1.0190, 1.0143]$ and $\beta_2 \in [1.7170, 4.9356]$. The given intervals contain the corresponding intervals of skewness and kurtosis coefficients of the SN model, which are $(-0.9953, 0.9953)$ and $[3.0000, 3.8692)$, respectively, and the PN model, which are $[-0.6115, 0.9007]$ and $[1.7170, 4.3556]$, respectively. More details can be found in Pewsey et al. [4]. The previous results illustrate the fact that the LGN model contains models with greater (and smaller) asymmetry degree than both the SN and PN models.

2.2. Distribution Function

In this section, we present the explicit formula for the CDF of LGN distribution.

Proposition 3. *Let $X \sim LGN(\delta, \gamma)$, then*

$$F_{LGN}(x) = \frac{\Gamma(\delta, -\gamma \ln \Phi(x))}{\Gamma(\delta)}, \tag{4}$$

where $\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}du$ is the gamma function and $\Gamma(a, x) = \int_x^\infty u^{a-1}e^{-u}dx$ is the upper incomplete gamma function.

Proof. The CDF of the LGN distribution is obtained as follows:

$$\begin{aligned}
F_{LGN}(x) &= \int_{-\infty}^{x} f_{LGN}(t)dt \\
&= \int_{-\infty}^{x} \frac{\gamma^\delta}{\Gamma(\delta)}[-\ln \Phi(t)]^{\delta-1}[\Phi(t)]^{\gamma-1}\phi(t)dt \\
&= -\int_{+\infty}^{-\ln \Phi(x)} \frac{\gamma^\delta}{\Gamma(\delta)} s^{\delta-1}e^{-\gamma s}ds; \text{ by } s = -\ln \Phi(t) \\
&= \int_{-\ln \Phi(x)}^{+\infty} \frac{\gamma^\delta}{\Gamma(\delta)} s^{\delta-1}e^{-\gamma s}ds \\
&= \frac{\Gamma\left(\delta, -\gamma \ln \Phi(x)\right)}{\Gamma(\delta)}
\end{aligned}$$

□

It can be shown that (see Cordeiro et al. [16]) for the density function given in (2), the quantile function is given by:

$$Q(u) = \Phi^{-1}\left(\exp\left\{-\gamma^{-1}Q^{-1}(\delta, u)\right\}\right),$$

where $Q^{-1}(\delta, u)$ is the inverse function of $Q(\delta, u) = \Gamma(\delta, u)/(\Gamma(\delta))$.

The inversion method can be used to generate a random variable with LGN distribution. Thus, let $\delta, \gamma \in \mathbb{R}^+$ and U be a random variable with uniform distribution, namely $U \sim \mathrm{U}(0,1)$. Then, the random variable X with distribution $\mathrm{LGN}(\delta, \gamma)$ can be obtained by letting

$$X = \Phi^{-1}\left(e^{-F^{-1}(1-U,\delta,\gamma)}\right),$$

where $\Phi^{-1}(\cdot)$ and $F^{-1}(\cdot, \delta, \gamma)$ are the inverses of the CDF of the normal and gamma distributions, respectively. The survival and hazard functions for the LGN distribution can be obtained from (2) and (4), and they are given by

$$S(t) = \Gamma(-\ln \Phi(t); \delta, \gamma),$$

and

$$r(t) = \frac{\frac{\gamma^\delta}{\Gamma(\delta)}[-\ln \Phi(t)]^{\delta-1}[\Phi(t)]^{\gamma-1}\phi(t)}{\Gamma(-\ln \Phi(t), \delta, \gamma)},$$

respectively.

2.3. Location-Scale Extension

Let $X \sim \mathrm{LGN}(\delta, \gamma)$. The location-scale extension of the random variable X is defined using the transformation $Y = \xi + \sigma X$, where $\xi \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$. The corresponding PDF of Y is given by

$$f_{LGN}(y; \xi, \sigma, \delta, \gamma) = \frac{\gamma^\delta}{\sigma\Gamma(\delta)}\left[-\ln \Phi\left(\frac{y-\xi}{\sigma}\right)\right]^{\delta-1}\left[\Phi\left(\frac{y-\xi}{\sigma}\right)\right]^{\gamma-1}\phi\left(\frac{y-\xi}{\sigma}\right), \quad (5)$$

where ξ is a location parameter and σ is a scale parameter. The random variable Y that has a distribution with density function given in Equation (5) is denoted as $Y \sim \mathrm{LGN}(\xi, \sigma, \delta, \gamma)$.

The previous representation of location scale can be extended to the case where response variable depends on regressor variables, say $Z_1, \ldots, Z_p$, through the relationship $\xi_i = \mathbf{z}_i^\top \boldsymbol{\beta}$; where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is an unknown vector of regression coefficients and $\mathbf{z} = (1, z_1, \ldots, z_p)^\top$ is a vector of known regressors correlated with the response vector.

The rth moment of a variable $Y \sim \mathrm{LGN}(\xi, \sigma, \delta, \gamma)$ can be obtained using the formula

$$\mathbb{E}[Y^r] = \sum_{l=0}^{r}\binom{r}{l}\xi^l\sigma^{r-l}\mathbb{E}[X^{r-l}], \quad r = 1, \ldots,$$

where $X \sim \mathrm{LGN}(\delta, \gamma)$.

Proof. Let $X \sim \mathrm{LGN}(\delta, \gamma)$, then, for $Y = \xi + \sigma X$ and $r = 1, \ldots$ it has

$$\begin{aligned}\mathbb{E}[Y^r] &= \mathbb{E}\left[(\xi + \sigma X)^r\right] \\ &= \mathbb{E}\left[\sum_{l=0}^{r}\binom{r}{l}\xi^l(\sigma X)^{r-l}\right] \\ &= \sum_{l=0}^{r}\binom{r}{l}\xi^l\sigma^{r-l}\mathbb{E}[X^{r-l}].\end{aligned}$$

In the second line, the binomial theorem is used. □

3. Log-Gamma-Normal Regression Model

Regression models have been a statistical technique widely used in many areas of knowledge to explain the behavior of a response variable, say Y, as a function of other variables called regressors, say $Z_1, \ldots, Z_p$, and a vector of unknown parameters called regression coefficients denoted by $\boldsymbol{\beta}$. Specifically, for a random sample of n individuals indexed by $i = 1, \ldots, n$, we have

$$y_i = \mathbf{z}_i^\top \boldsymbol{\beta} + \varepsilon_i \quad \text{for} \quad i = 1, 2, \ldots, n; \tag{6}$$

where ε_i is a random variable (random error) with certain PDF, the most common being the normal distribution assumption, i.e., $\varepsilon_i \sim N(0, \sigma^2)$. Given the multiple departures from the normality assumption and the actual behavior of the random variable ε_i, this assumption has been replaced in numerous instances by other more realistic ones, usually looking for distributions to fit data with higher or lower skewness and/or kurtosis than that allowed by the normal distribution. Notable inferential mistakes are made (invalid results) when we work under the normal assumption and this assumption is not true. In some cases, a simple transformation helps to solve this problem, but this strategy typically has problems of interpretability of the results or the coefficients of the model.

Now, we change the normal assumption using the LGN assumption in the random error term ε_i, so we suppose that $\varepsilon_i \sim \text{LGN}(0, \sigma, \delta, \gamma)$ and this leads to $Y_i \sim \text{LGN}(\mathbf{z}_i^\top \boldsymbol{\beta}, \sigma, \delta, \gamma)$ for $i = 1, \ldots, n$. The case $\delta = \gamma = 1$ follows the ordinary normal regression model. Using the least squares method, we obtain the estimators $\tilde{\boldsymbol{\beta}} = \left(\mathbf{z}^\top \mathbf{z}\right)^{-1} \mathbf{z}^\top \mathbf{y}$, which are not unbiased for the parameters of the regression coefficients but the correction $\tilde{\beta}_0^* = \tilde{\beta}_0 + \hat{\mathbb{E}}[\varepsilon]$, where the last term represents the estimated expected value of the random variable ε, such that we can obtain unbiased estimators of the parameters.

Estimation Using Maximum Likelihood Method

We initially define some quantities: $\mathbf{Z}$ is a matrix $n \times (p+1)$ where rows $\mathbf{z}_i$ correspond to observations for the ith individual for p independent variables; $\mathbf{y}$ is a vector $n \times 1$ corresponding to responses for the ith individual; and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is an unknown vector of regression coefficients. Thus, given a random sample of size n, say $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$, where $Y_i \sim \text{LGN}(\mathbf{z}_i^\top \boldsymbol{\beta}; \sigma, \delta, \gamma)$ for $i = 1, \ldots, n$; the log-likelihood function for the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma, \delta, \gamma)^\top$ can be written as follows:

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{y}) = {} & n[\delta \ln(\gamma) - \ln(\Gamma(\delta)) - \ln(\sigma) - 0.5 \ln(2\pi)] - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) \\ & + (\delta - 1) \sum_{i=1}^{n} \ln \left[-\ln \Phi \left(\frac{y_i - \mathbf{z}_i^\top \boldsymbol{\beta}}{\sigma} \right) \right] + (\gamma - 1) \sum_{i=1}^{n} \ln \left[\Phi \left(\frac{y_i - \mathbf{z}_i^\top \boldsymbol{\beta}}{\sigma} \right) \right]. \end{aligned} \tag{7}$$

After taking the first partial derivatives of the log-likelihood function (7) regarding the parameters of interest and setting them equal to zero, we obtain the following score equations:

$$U(\boldsymbol{\beta}) = \frac{1}{\sigma^2} \mathbf{Z}^\top (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) - \frac{1}{\sigma} \mathbf{Z}^\top [(\delta - 1)\mathbf{U} + (\gamma - 1)\mathbf{I}_n] \boldsymbol{\Delta}_1 = 0, \tag{8}$$

$$\begin{aligned} U(\sigma) = {} & -\frac{n}{\sigma} + \frac{1}{\sigma^3} (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) \\ & - \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})^\top [(\delta - 1)\mathbf{U} + (\gamma - 1)\mathbf{I}_n] \boldsymbol{\Delta}_1 = 0, \end{aligned} \tag{9}$$

$$U(\delta) = n \ln(\gamma) - n\psi(\delta) + \sum_{i=1}^{n} \ln[-\ln \Phi(x_i)] = 0, \tag{10}$$

$$U(\gamma) = \frac{n\delta}{\gamma} + \sum_{i=1}^{n} \ln[\Phi(x_i)] = 0, \tag{11}$$

where $\boldsymbol{\Delta}_1 = (v_1, \ldots, v_n)^\top$, and $\mathbf{U} = \text{diag}\{1/u_1, \ldots, 1/u_n\}$ with $v_i = \phi(x_i)/\Phi(x_i)$ and $u_i = \ln[\Phi(x_i)]$; and $x_i = (y_i - \mathbf{z}_i^\top \boldsymbol{\beta})/\sigma$ for $i = 1, \ldots, n$; $\mathbf{I}_n$ is the identity matrix of order n, and $\psi(\cdot)$ is the digamma function.

The elements of the observed information matrix for the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma, \delta, \gamma)^\top$ are easily computed by taking second partial derivatives, obtaining:

$$
\begin{aligned}
j_{\boldsymbol{\beta}^\top} &= \frac{1}{\sigma^2}\mathbf{Z}^\top\mathbf{Z} + \frac{\delta-1}{\sigma^2}\mathbf{Z}^\top\boldsymbol{\Delta}_2\mathbf{Z} + \frac{\gamma-1}{\sigma^2}\mathbf{Z}^\top\boldsymbol{\Delta}_3\mathbf{Z},\\
j_{\boldsymbol{\beta}\sigma} &= \frac{2}{\sigma^3}\mathbf{Z}^\top(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta}) + \frac{\delta-1}{\sigma^2}\mathbf{Z}^\top\boldsymbol{\Delta}_4 + \frac{\gamma-1}{\sigma^2}\mathbf{Z}^\top\boldsymbol{\Delta}_5,\\
j_{\sigma\sigma} &= -\frac{n}{\sigma^2} + \frac{3}{\sigma^2}\sum_{i=1}^{n}x_i^2 + \frac{\delta-1}{\sigma^2}\sum_{i=1}^{n}\left\{-2\frac{v_i}{u_i}x_i + \frac{v_i^2}{u_i^2}x_i^2 + \frac{v_i x_i^3 + v_i^2 x_i^2}{u_i}\right\}\\
&\quad + \frac{\gamma-1}{\sigma^2}\sum_{i=1}^{n}\left[-2v_i x_i + v_i x_i^3 + x_i^2 v_i^2\right],\\
j_{\boldsymbol{\beta}\gamma} &= \frac{1}{\sigma}\mathbf{Z}^\top\boldsymbol{\Delta}_1, \qquad j_{\boldsymbol{\beta}\delta} = \frac{1}{\sigma}\mathbf{Z}^\top\boldsymbol{\Delta}_6, \qquad j_{\sigma\delta} = \frac{1}{\sigma}\mathbf{X}^\top\boldsymbol{\Delta}_6, \qquad j_{\sigma\gamma} = \frac{1}{\sigma}\mathbf{X}^\top\boldsymbol{\Delta}_1,\\
j_{\delta\delta} &= n\psi_1(\delta), \qquad j_{\delta\gamma} = -\frac{n}{\gamma}, \qquad j_{\gamma\gamma} = \frac{n\delta}{\gamma^2},
\end{aligned}
$$

where $\boldsymbol{\Delta}_2 = \operatorname{diag}\{v_i^2/u_i^2 + (v_i x_i + v_i^2)/u_i\}$ and $\boldsymbol{\Delta}_3 = \operatorname{diag}\{v_i x_i + v_i^2\}$ with $i = 1, \ldots, n$; $\boldsymbol{\Delta}_4 = (a_1, \ldots, a_n)^\top$, $\boldsymbol{\Delta}_5 = (b_1, \ldots, b_n)^\top$, and $\boldsymbol{\Delta}_6 = (c_1, \ldots, c_n)^\top$ with $a_i = -v_i/u_i + v_i^2 x_i/u_i^2 + (v_i x_i^2 + v_i^2 x_i)/u_i$, $b_i = v_i x_i^2 + v_i^2 x_i - v_i$, $c_i = v_i/u_i$ for $i = 1, \ldots, n$, $\psi_1(\cdot)$ is the trigamma function and $\mathbf{X} = (x_1, \cdots, x_n)^\top$ with $x_i = (y_i - \mathbf{z}_i^\top\boldsymbol{\beta})/\sigma$ for $i = 1, \ldots, n$.

The Fisher information matrix $I(\boldsymbol{\theta})$ can be obtained numerically, calculating n^{-1} times the expected value of the observed information matrix. When $\delta = \gamma = 1$, we obtain the case of the normal distribution $N(0, \sigma^2)$ for the random variable ε_i. Using numerical approximation, the determinant of the Fisher information matrix is $\det(I(\boldsymbol{\theta})) = \det(\mathbf{Z}^\top\mathbf{Z})[-0.3137\det(\mathbf{Z}^\top\mathbf{Z}) + 0.3093\sum_{j=0}^{p}\bar{z}_j^2]$, where $\det(\cdot)$ denotes the determinant function of a matrix and $\bar{z}_j$ denotes the mean in the sample of the variable Z_j. Thus, the determinant of the information matrix is different to zero, and the information matrix is non-singular, ensuring the conditions to apply asymptotic approximation to the normal distribution of the maximum likelihood estimator vector of $\boldsymbol{\theta}$. Here, the covariances matrix of $\hat{\boldsymbol{\theta}}$ is the inverse of the Fisher information matrix, i.e., $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} = I^{-1}(\boldsymbol{\theta})$.

Approximation $N_{p+4}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}})$ can be used to construct confidence intervals for $\boldsymbol{\theta}_r$, which are given by $\hat{\boldsymbol{\theta}}_r \mp z_{1-\alpha/2}\sqrt{\hat{\sigma}(\hat{\boldsymbol{\theta}}_r)}$, where $\hat{\sigma}(\hat{\boldsymbol{\theta}}_r)$ corresponds to the rth diagonal element of the matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$ and $z_{1-\alpha/2}$ denotes $100(1-\alpha/2)$ quantile of the standard normal distribution.

4. Censored LGN Model

Models for censored data are common in economic research, medicine, biology, and survival analysis. Usually, this type of data is analyzed using the Tobit model (see Tobin [24], also known as censored normal model (CN)). In some cases, the tails of the distribution of the random errors are more or less heavy than the tails of the normal distribution, consequently showing that the Tobit model does not estimate the probability in the censored part very well, and this leads to bad estimates. In these cases, it must be assumed that another distribution to model errors, especially in the case of asymmetric errors, can work with the power-normal Tobit model (PNT) (see Martínez-Flórez et al. [25]), the censored SN model, or any other model that fits the degree of asymmetry and the kurtosis of the errors in the model. We now extend the LGN regression model to the censored data, which we will call the censored LGN regression model (CLGN).

Censored LGN Variable

Consider a random variable $Y^* \sim \text{LGN}(\xi, \sigma, \delta, \gamma)$ and let $\{y_1^*, y_2^*, \ldots, y_n^*\}$ be a random sample of size n of Y^*. Let T be a value of censorship for the Y^* variable. The CLGN random variable Y is defined as

$$y_i = \begin{cases} y_i^*, & \text{if } y_i^* > T, \\ T, & \text{if } y_i^* \leq T, \end{cases}$$

for $i = 1, \ldots, n$. We use the notation CLGN$(\xi, \sigma, \delta, \gamma)$. Consequently, the probability mass at the value T is $\Pr(y_i = T) = \Pr(y_i^* \leq T) = 1 - \Gamma(-\ln \Phi(z_{Ti}); \delta, \gamma)$, where $z_{Ti} = (T - \xi)/\sigma$. For $y_i^* > T$, the distribution of the variable Y is LGN$(\xi, \sigma, \delta, \gamma)$. Although the formulation above the threshold T is not null, it can be transformed back to zero by taking $y_i^* - T$. Hence, there is no loss of generality in taking $T = 0$.

When we have regressor variables, say $Z_1, \ldots, Z_p$, through the relationship $\xi_i = \mathbf{z}_i^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is an unknown vector of regression coefficients, and $\mathbf{z} = (1, z_1, \ldots, z_p)^\top$ is a vector of known regressors correlated with the response vector, we have a CLGN regression model defined by the random variable $y_i = \max\{y_i^*, T\}$, with $y_i^* = \mathbf{z}_i^\top \boldsymbol{\beta} + \varepsilon_i$, $i = 1, \ldots, n$; i.e.,

$$y_i = \begin{cases} \mathbf{z}_i^\top \boldsymbol{\beta} + \varepsilon_i, & \text{if } \mathbf{z}_i^\top \boldsymbol{\beta} + \varepsilon_i > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

For a sample of size n, $\mathbf{y} = (y_1, \ldots, y_n)^\top$, where $Y_i \sim$ CLGN$(\mathbf{z}_i^\top \boldsymbol{\beta}; \sigma, \delta, \gamma)$ for $i = 1, \ldots, n$; the log-likelihood function for the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma, \delta, \gamma)^\top$ is given by

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = \sum\nolimits_0 [1 - \Gamma(-\ln \Phi(x_{Ti}); \delta, \gamma)] + \sum\nolimits_1 \left[\ln\left(\frac{\gamma^\delta}{\sigma \Gamma(\delta)} \right) + (\delta - 1) \ln[-\ln \Phi(x_i)] \right] + \sum\nolimits_1 \left[-\frac{1}{2\sigma^2} x_i^2 + (\gamma - 1) \ln[\Phi(x_i)] \right],$$

where $\sum_0$ and $\sum_1$ denote the sum in the censored part and uncensored part, respectively; $x_i = (y_i - \mathbf{z}_i^\top \boldsymbol{\beta})/\sigma$ and $x_{Ti} = (T - \mathbf{z}_i^\top \boldsymbol{\beta})/\sigma$.

Special cases from model (12) occur when $\delta = \gamma = 1$, so the Tobit model follows (see Tobin [24]) and with $\delta = 1$ the Tobit PN model follows (see Martínez-Flórez et al. [25]). The parameters estimation can be performed by the maximum likelihood method, i.e., by maximizing the function $\ell(\boldsymbol{\theta}; \mathbf{y})$, whose solution using iterative numerical methods leads to the maximum likelihood estimator (MLE) of the model.

5. Simulation Study

To study the performance of the MLE $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\sigma}, \hat{\delta}, \hat{\gamma})^\top$ of parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma, \delta, \gamma)^\top$, we conducted a Monte Carlo simulation study with small and moderate samples. In the study, we generated 5000 samples of sizes $n = 50$, 100, 200 and 500, and we considered the LGN model. The following parameter values were taken: $\delta, \gamma = 0.75, 1.50$; $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top = (2.0, 1.0)^\top$ and we took $\sigma = 0.50$.

We considered a linear model with a single covariate Z whose values were generated according to a uniform distribution U$(0, 1)$. We also took errors $\varepsilon_i \sim$ LGN$(0, \sigma, \delta, \gamma)$. To evaluate estimators performance for point estimates we considered the bias (Bias), the relative bias (RB) defined as (absolute value of bias / true parameter value) and the square root of the mean squared error $RMSE = \sqrt{\text{MSE}}$, which is the mean over all samples of the squared bias plus the variance. Maximum likelihood parameter estimates were computed using the optim function in statistical package R Development Core Team [23].

Tables 1 and 2 present the results of the simulation study. It can be seen from the table that the RMSEs of MLEs for β_0, β_1, σ, δ and γ decreases as sample sizes increase, which is expected since estimators are consistent. The relative bias of the MLEs also decrease as sample sizes increase. The MLEs of β_0 are unstable because this parameter is affected by the asymmetry parameter; however, its MLE becomes more stable as the sample size becomes larger. It can also be seen that when the parameter γ increases, the bias of the

MLEs of the β_0, δ and γ is larger. The main conclusion is that we are quite safe working with the MLEs if sample sizes are greater than 100.

Table 1. Performance evaluation for the MLE of β_0, β_1, σ, δ and γ under LGN model for $\delta = 0.75$ and $\gamma = 0.75, 1.50$.

		$\delta = 0.75, \gamma = 0.75$			$\delta = 0.75, \gamma = 1.50$		
n	$\hat{\theta}$	**Bias**	**RB**	**RMSE**	**Bias**	**RB**	**RMSE**
50	$\hat{\beta}_0$	0.027	1.3	0.625	0.093	4.6	0.605
	$\hat{\beta}_1$	0.003	0.3	0.207	0.003	0.3	0.171
	$\hat{\sigma}$	−0.034	6.8	0.194	−0.049	9.8	0.193
	$\hat{\delta}$	0.124	16.6	0.586	0.243	32.4	1.187
	$\hat{\gamma}$	0.439	58.5	1.898	0.464	30.9	2.304
100	$\hat{\beta}_0$	0.024	1.2	0.489	0.047	2.4	0.455
	$\hat{\beta}_1$	0.001	0.1	0.143	0.001	0.1	0.119
	$\hat{\sigma}$	−0.021	4.2	0.166	−0.019	3.8	0.165
	$\hat{\delta}$	0.112	15.0	0.526	0.208	27.7	0.953
	$\hat{\gamma}$	0.291	38.8	1.242	0.414	27.6	1.946
200	$\hat{\beta}_0$	0.010	0.5	0.344	0.020	1.0	0.308
	$\hat{\beta}_1$	0.001	0.1	0.102	0.001	0.1	0.085
	$\hat{\sigma}$	−0.018	3.7	0.142	−0.017	3.4	0.155
	$\hat{\delta}$	0.111	14.8	0.450	0.162	21.6	0.760
	$\hat{\gamma}$	0.214	28.5	0.908	0.393	26.2	1.639
500	$\hat{\beta}_0$	0.004	0.2	0.163	0.006	0.3	0.157
	$\hat{\beta}_1$	0.001	0.1	0.051	0.001	0.0	0.042
	$\hat{\sigma}$	−0.004	0.3	0.108	−0.006	1.2	0.126
	$\hat{\delta}$	0.072	9.6	0.321	0.085	11.3	0.430
	$\hat{\gamma}$	0.124	16.5	0.543	0.281	18.7	1.108

Table 2. Performance evaluation for the MLE of β_0, β_1, σ, δ and γ under LGN model for $\delta = 1.50$ and $\gamma = 0.75, 1.50$.

		$\delta = 1.50, \gamma = 0.75$			$\delta = 1.50, \gamma = 1.50$		
N	$\hat{\theta}$	**Bias**	**RB**	**RMSE**	**Bias**	**RB**	**RMSE**
50	$\hat{\beta}_0$	0.018	0.9	0.673	0.095	4.7	0.660
	$\hat{\beta}_1$	−0.003	0.3	0.180	−0.002	0.2	0.144
	$\hat{\sigma}$	−0.032	6.3	0.273	−0.069	13.9	0.188
	$\hat{\delta}$	0.316	21.1	1.747	0.269	17.9	1.364
	$\hat{\gamma}$	0.593	79.5	2.210	0.422	28.2	2.355
100	$\hat{\beta}_0$	0.013	0.6	0.522	0.058	2.9	0.540
	$\hat{\beta}_1$	−0.001	0.1	0.126	−0.001	0.1	0.100
	$\hat{\sigma}$	−0.017	3.3	0.244	−0.026	5.2	0.159
	$\hat{\delta}$	0.308	20.6	1.527	0.257	17.1	1.180
	$\hat{\gamma}$	0.446	59.5	1.647	0.415	27.7	2.016
200	$\hat{\beta}_0$	0.012	0.6	0.387	0.034	1.7	0.406
	$\hat{\beta}_1$	−0.001	0.1	0.090	−0.001	0.1	0.072
	$\hat{\sigma}$	−0.014	2.8	0.198	−0.014	2.7	0.135
	$\hat{\delta}$	0.246	16.4	1.222	0.231	15.4	0.991
	$\hat{\gamma}$	0.309	41.2	1.281	0.344	22.9	1.602
500	$\hat{\beta}_0$	0.011	0.5	0.212	0.001	0.1	0.213
	$\hat{\beta}_1$	−0.001	0.1	0.045	−0.001	0.1	0.036
	$\hat{\sigma}$	−0.003	0.6	0.124	−0.001	0.1	0.102
	$\hat{\delta}$	0.150	10.0	0.757	0.131	8.7	0.628
	$\hat{\gamma}$	0.139	18.6	0.657	0.240	16.0	1.043

6. Real Data Applications

6.1. Application 1

We consider a dataset related to longitudinal data on cholesterol levels collected as part of the famed Framingham heart study. The file includes information for $n = 200$ randomly selected individuals, reported in Zhang and Davidian [22]. The considered variables were the cholesterol level (Y), the age of the individual at baseline (Z_1) and the gender indicator (Z_2) (0 = female, 1 = male). For this application, we take only the observations in the second period of time of the measurement $(n = 176)$. Table 3 presents the summary statistic, including measures of skewness and kurtosis for cholesterol data. Clearly, the values of the skewness and kurtosis for cholesterol data justify using an asymmetric model, the PN, SN or LGN model.

Table 3. Summary statistics for cholesterol levels for 176 subjects of the Framingham cholesterol study.

Mean	SD	$\sqrt{b_1}$	b_2
224.597	41.242	0.896	3.594

A model with errors following a normal distribution was fitted, and it was found that the Shapiro–Wilk normality test gives a value of the test statistic $W = 0.9599$ with p-value = 6.254×10^{-5}, so the normality of the errors is rejected. We fitted linear regression models by assuming errors following an asymmetric distribution, namely SN, PN and LGN distributions. For estimating parameters in the considered models, we use the optim function available in R Development Core Team [23].

Table 4 presents the MLE for the estimated parameters of the fitted models. We took the obtained estimates from the normal model using the function lm R Development Core Team [23] as the initial values. For δ and γ (and some cases σ) we took the obtained estimates under the SN, PN and LGN location-scale models fitted to Y variable. From the table, the age at baseline variable (Z_1) is not significant and the cholesterol level depends solely on the gender in normal, PN and SN models. For the LGN model, it follows that the cholesterol level depends on the sex and the age of the individual at the baseline.

The considered linear model was

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \varepsilon_i, \quad i = 1, 2, \ldots, 176.$$

Table 4. Estimates and standard error (SE) for normal, PN, SN and LGN linear regression models fitted to cholesterol data.

	Normal	PN	SN	LGN
β_0	150.6 (15.9)	71.1 (25.0)	128.1 (14.0)	90.9 (0.2)
β_1	−9.3 (5.8)	−7.7 (5.5)	−5.7 (5.0)	−5.3 (2.6)
β_2	1.9 (0.4)	1.6 (0.4)	1.3 (0.4)	1.1 (0.2)
σ	38.4 (2.1)	61.8 (6.1)	60.3 (4.8)	29.2 (0.1)
λ	-	-	3.6 (1.1)	-
δ	-	-	-	0.2 (0.1)
γ	-	8.2 (3.5)	-	9.2 (4.1)

To compare the normal, PN and LGN models, which are nested models, we used the AIC, by Akaike [26], AICc (corrected Akaike information criterion), and BIC (Bayesian information criterion) by Schwarz [27], which are written as

$$\begin{aligned} AIC &= -2\hat{\ell}(\cdot) + 2k, \\ AICc &= AIC + \big(2k(k+1)\big)/\big(n - (k+1)\big), \\ BIC &= -2\hat{\ell}(\cdot) + k\log(n), \end{aligned}$$

where k is the number of unknown parameters in the considered model. The best model is

the one with the smallest AIC or AICc or BIC.

Using the Normal, PN, SN and LGN distributions, the scaled residuals $e_i = (y_i - \mathbf{z}_i^\top \hat{\boldsymbol{\beta}})/\hat{\sigma}$ are evaluated and presented in Figures 2 and 3.

The normality assumption for errors can be tested by the hypothesis

$$H_{01} : (\delta, \gamma) = (1, 1) \quad \text{versus} \quad H_{11} : (\delta, \gamma) \neq (1, 1),$$

using the likelihood-ratio (LR) statistics, $-2\log(\Lambda_1) = -2\big(\ell_N(\hat{\boldsymbol{\theta}}) - \ell_{LGN}(\hat{\boldsymbol{\theta}})\big)$, which for the dataset under study, leads to $-2\log(\Lambda_1) = 18.228$, so that $p-$value< 0.05, with strong indication against the null hypothesis.

Similarly, the assumption of PN distribution for the errors can be tested by the hypothesis

$$H_{02} : \delta = 1 \quad \text{versus} \quad H_{12} : \delta \neq 1,$$

using the LR statistics, $-2\log(\Lambda_2) = -2\big(\ell_{PN}(\hat{\boldsymbol{\theta}}) - \ell_{LGN}(\hat{\boldsymbol{\theta}})\big)$, which leads to $-2\log(\Lambda_2) = 6.622$, so that p-value < 0.05, with strong indication against the null hypothesis.

Table 5 presents the AIC, AICc and BIC criteria for the normal, PN, SN and LGN models. Please note that according to these criteria, the model that best fits the dataset is the SN, since it has a lower value of AIC, AICc and BIC, followed by the LGN model. However, we remember that the SN model presents a singular information matrix when the asymmetry parameter λ is zero, and therefore, hypothesis tests about the model parameters using likelihood-ratio statistics are not feasible from the theory of large samples; for example, for testing the significance of the asymmetry parameter in the SN model. This constitutes a disadvantage related to the LGN model, for which it was shown that it has a non-singular information matrix. In addition, as mentioned in Section 2, the LGN model has higher ranges of asymmetry and kurtosis than the SN model, so in practice it may be preferable in certain situations.

Table 5. AIC, AICc, and BIC for normal, PN and LGN linear models.

Criteria	Normal Model	PN Model	SN Model	LGN Model
AIC	1789.584	1779.979	1773.770	1775.356
AICc	1789.817	1780.331	1764.123	1775.853
BIC	1799.096	1795.832	1789.622	1794.379

This discussion illustrates that the final selection of a model is often simply a matter of choice. The LGN model can be considered appropriate if we want to use a model with which we can carry out hypothesis tests about the parameters, especially those associated with skewness and kurtosis in the model. In any case, the final choice must be duly justified.

For non-nested models, we used a generalized LR statistic test studied by Vuong [28]. This test was derived to compare competing models that are strictly non-nested. Since F_θ and G_ζ are two non-nested models, $f(y_i \mid x_i, \theta)$ and $g(y_i \mid x_i, \zeta)$ two densities corresponding to these non-nested models, the LR statistics to compare both models is given by

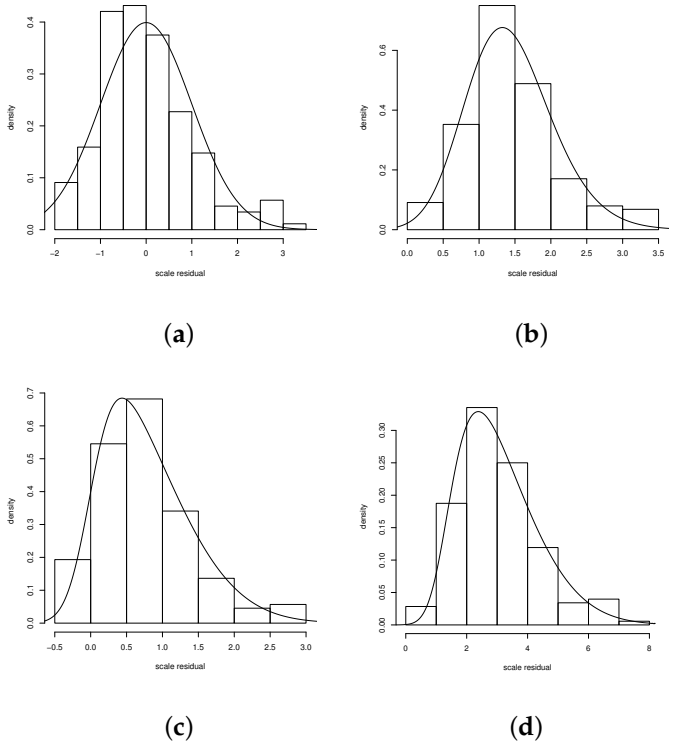
$$LR(\hat{\theta}, \hat{\zeta}) = \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \log \frac{f(y_i \mid x_i, \hat{\theta})}{g(y_i \mid x_i, \hat{\zeta})} \right\},$$

which does not follow a chi-square distribution. To overcome this problem, Vuong [28] proposed an alternative approach based on the Kullback–Liebler information criterion [29]. Based on the distance between each model and the true process generating the data, namely the model $h^0(y \mid x)$, he arrived at the statistics

$$T_{LR,NN} = \frac{1}{\sqrt{n}} \frac{LR(\hat{\theta}, \hat{\zeta})}{\hat{w}}, \tag{13}$$

where

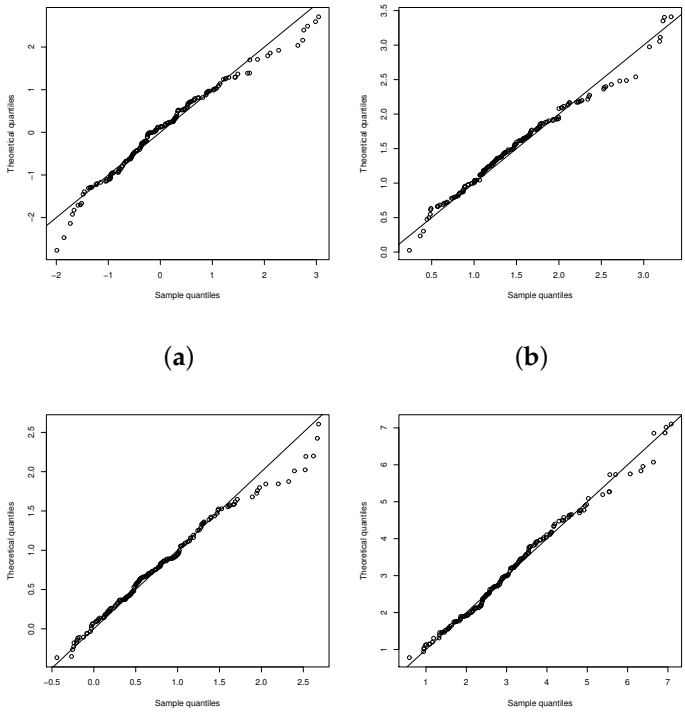
$$\hat{w}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{f(y_i \mid x_i,\hat{\theta})}{g(y_i \mid x_i,\hat{\zeta})}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(y_i \mid x_i,\hat{\theta})}{g(y_i \mid x_i,\hat{\zeta})}\right)^2.$$



(**a**) (**b**)

(**c**) (**d**)

Figure 2. Histogram for scaled residuals for (**a**) Normal model, (**b**) PN model, (**c**) SN model, and (**d**) LGN model fitted to the cholesterol data.



(**a**) (**b**)

(**c**) (**d**)

Figure 3. QQplots for (**a**) Normal model, (**b**) PN model, (**c**) SN model, and (**d**) LGN model fitted to the cholesterol data.

For strictly non-nested models, the statistic (13) converges in distribution to a standard normal distribution under the null hypothesis of equivalence of the models. Thus, the null hypothesis is not rejected if $|T_{LR,NN}| \leq z_{p/2}$. On the other hand, we reject at significance level p the null hypothesis of equivalence of the models in favor of model F_θ being better (or worse) than model G_ζ if $T_{LR,NN} > z_p$ (or $T_{LR,NN} < -z_p$).

We now use Voung for comparing the LGN versus SN and PN versus SN models fitted to the data, since they are two non-nested models. Let $f\left(y_i \mid x_i, \hat{\theta}\right)$ be the LGN model and $g\left(y_i \mid x_i, \hat{\zeta}\right)$ the SN model. The generalized LR test statistic value is $T_{LR,NN} = 33.981$. For PN versus SN, the generalized LR test statistic value is $T_{LR,NN} = 30.072$. Therefore, the LGN and PN models are significantly superior to the SN model, according to the generalized LR statistic. Then, the LGN model is the better model compared with the normal, PN and SN models.

6.2. Application 2

For the second application, we consider a dataset consisting of measurements for 68 solar-type stars. These data were previously described and analyzed by Santos et al. [30] and Tovar-Falón et al. [31]. The dataset is available in the astrodatR library of the R Development Core Team [23] package under the name Stellar Abundances. In this application, we consider the response variable: $\log N(Be)$, which represents the log of the abundance of beryllium scaled to Sun's abundance, i.e., the Sun has $\log N(Be) = 0.0$. The explanatory variable is Teff/1000, which represents the effective stellar surface temperature (in Kelvin).

In astronomy, objects such as stars, galaxies or X-ray sources, among others, are observed in some new wave bands. Some of these objects can go unnoticed due to limited sensitivities, leading to upper limits in the measurement of their luminosity (see Feilgelson [32]). For the dataset, 14 observations (19.35%) were censored at 0.0, i.e., 12 beryllium measurements were not detected.

We fitted the censored normal (CN) or Tobit model using the censReg function of R Development Core Team [23]. Likewise, we also fitted the censored power-normal (CPN) and censored LGN (CLGN) models. The Table 6 shows the MLEs of the fitted models. The initial values for the parameters β_k were initially taken from those returned by the censReg package of the CN model. The outputs show that the explanatory variable X is significant in the considered models.

Table 6. Estimates (standard error) for CN, CPN, and CLGN linear models.

Parameters	CN Model	CPN Model	CLGN Model
β_0	−0.9450 (0.5854)	−1.6054 (0.6910)	−1.1897 (0.4331)
β_1	0.3224 (0.1023)	0.5222 (0.1208)	0.3520 (0.0732)
σ	0.3147 (0.0281)	0.0813 (0.0688)	0.3268 (0.0326)
α	-	-	2.0123 (0.5762)
λ	-	0.0280 (0.0477)	4.5803 (0.9748)

Table 7 contains the AIC and AICC values for the fitted models, where it is observed that the CLGN model presents the best fit. Figure 4a–c show the histogram, the CDF and the qqplot of the CLGN model of the scale residual errors of the uncensored part. Here one can see the good fit of the CLGN model.

Table 7. AI and AICC for CN, CPN and CLGN linear models.

Criteria	CN Model	CPN Model	CLGN Model
AIC	50.2585	63.9628	−25.8152
AICC	50.6335	64.5977	−24.8474
BIC	62.9170	72.8408	−14.7176

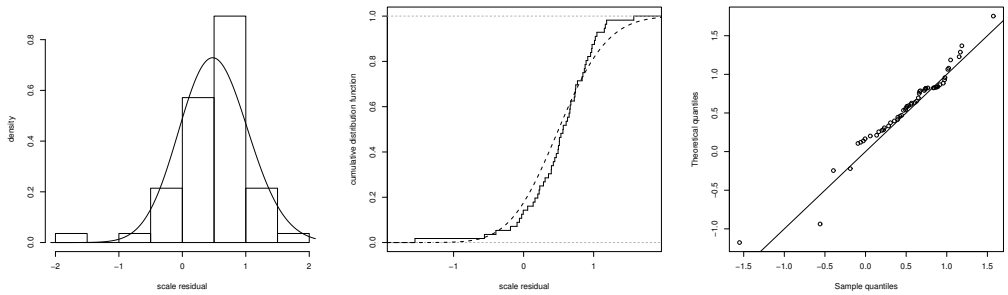
We compare the Normal and PN models against the LGN model, so for hypothesis testing

$$(\delta, \gamma) = (1,1) \quad \text{versus} \quad (\delta, \gamma) \neq (1,1)$$

and

$$\delta = 1 \quad \text{versus} \quad \delta \neq 1,$$

we have $-2\log(\Delta_1) = 8.4432$ and $-2\log(\Delta_2) = 20.1476$ both statistics with p-value < 0.05 for which both tests are rejected and therefore the LGN model performs better than the Normal and PN models.



(**a**) (**b**) (**c**)

Figure 4. (**a**) histogram for scaled residuals CLGN model, (**b**) CDF of the scaled residuals CLGN model (**c**) qqplot for scaled residuals CLGN regression model.

7. Conclusions

In this paper, we have proposed the asymmetric LGN distribution to give flexibility to the term of error in linear regression models. The LGN is based on the log-gamma-generated families of distributions of Amini et al. [15]. This new model presents greater ranges of asymmetry and kurtosis, and it extends the PN family of distribution; therefore, it has more flexibility in terms of asymmetry and kurtosis. The ordinary Tobit model Tobin [24] and the Tobit power-normal model Martínez-Flórez et al. [25] are special cases from an extension of the studied model LGN to the case of censored data. The maximum likelihood method was implemented, and the Fisher information matrix was derived, and it was shown numerically to be non-singular, which guarantees valid large sample results for the likelihood-ratio statistics. Two illustrations of real data reveal that the proposed model can be a useful alternative to existing models such as normal, power-normal, Tobit normal and Tobit power-normal. In addition, under certain considerations such as the non-singularity of the information matrix of the model and larger ranges of asymmetry and kurtosis, it may be a better alternative to the skew-normal distribution.

Author Contributions: Conceptualization, R.T.-F. and G.M.-F.; Methodology, R.T.-F., G.M.-F. and H.B.; Data curation, G.M.-F.; Formal analysis, R.T.-F., G.M.-F. and H.B.; Investigation, R.T.-F., G.M.-F. and H.B.; Resources, R.T.-F. and G.M.-F.; Software, R.T.-F. and G.M.-F.; Supervision, H.B.; Validation, G.M.-F. and R.T.-F.; Visualization, R.T.-F. and G.M.-F.;Writing—original draft, R.T.-F., G.M.-F. and H.B. and Writing—review and editing, R.T.-F., G.M.-F. and H.B. All authors have read and agreed to the published version of the manuscript.

Funding: Resolución de Problemas de Situaciones Reales Usando Análisis Estadístico a través del Modelamiento Multidimensional de Tasas y Proporciones; Esquemas de Monitoreamiento para Datos Asimétricos no Normales y una Estrategia Didáctica para el Desarrollo del Pensamiento Lógico-Matemático. Universidad de Córdoba, Colombia, Code FCB-05-19.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Details about data available are given in Section 6.

Acknowledgments: G. Martínez-Flórez and R. Tovar-Falón acknowledges the support given by Universidad de Córdoba, Montería, Colombia.

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

References

1. Azzalini, A. A class of distributions which includes the normal ones. *Scand. J. Stat.* **1985**, *12*, 171–178.
2. Durrans, S.R. Distributions of fractional order statistics in hydrology. *Water Resour. Res.* **1992**, *28*, 1649–1655. [CrossRef]
3. Gupta, R.D.; Gupta, R.C. Analyzing skewed data by power normal model. *Test* **2008**, *17*, 197–210. [CrossRef]
4. Pewsey, A.; Gómez, H.W.; Bolfarine, H. Likelihood-based inference for power distributions. *Test* **2012**, *21*, 775—789 [CrossRef]
5. Martínez-Flórez, G.; Bolfarine, H.; Gómez, H.W. Skew-normal alpha-power model. *Statistics* **2014**, *48*, 1414–1428. [CrossRef]
6. Martínez-Flórez, G.; Vergara-Cardozo, S.; González, L.M. The family of log-skew-normal alpha-power distributions using precipitation data. *Rev. Colomb. Estad.* **2013**, *36*, 43–57.
7. Tovar-Falón, R.; Bolfarine, H.; Martínez-Flórez, G. The Asymmetric Alpha-Power Skew-t Distribution. *Symmetry* **2020**, *12*, 82. [CrossRef]
8. Azzalini, A.; Capitanio, A. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-*t* distribution. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2003**, *65*, 367–389. [CrossRef]
9. Zhao, J.; Kim, H.M. Power-*t* distributions. *Commun. Stat. Appl. Methods* **2016**, *23*, 321–334.
10. Tung, H.P.; Tseng, S.T.; Hsu, N.J.; Hou, Y.T. A generalized pH acceleration model of nano-sol products and the effects of model misspecification on shelf-life prediction. *IISE Trans.* **2022**, *54*, 496–504. [CrossRef]
11. Martínez-Flórez, G.; Tovar-Falón, R.; Jimémez-Narváez, M. Likelihood-Based Inference for the Asymmetric Beta-Skew Alpha-Power Distribution. *Symmetry* **2020**, *12*, 613. [CrossRef]
12. Martínez-Flórez, G.; Tovar-Falón, R.; Martínez-Guerra, M. The Censored Beta-Skew Alpha-Power Distribution. *Symmetry* **2021**, *13*, 1114. [CrossRef]
13. Sahu, S.K.; Dey, D.K.; Branco, M.D. A new class of multivariate skew distributions with applications to Bayesian regression models. *Can. J. Stat.* **2003**, *31*, 129–150. [CrossRef]
14. Martínez-Flórez, G.; Bolfarine, H.; Gómez, H.W. Asymmetric regression models with limited responses with an application to antibody response to vaccine. *Biom. J.* **2013**, *55*, 156–172. [CrossRef]
15. Amini, M.; MirMostafaee, S.M.T.K.; Ahmani, J. Log–gamma–generated families of distributions. *Statistics* **2014**, *48*, 913–926. [CrossRef]
16. Cordeiro, G.M.; Bourguignon, M.; Ortega, E.M.M.; Ramires, T.G. General mathematical properties, regression and applications of the log-gamma-generated family. *Commun. Stat.—Theory Methods* **2018**, *47*, 1050–1070. [CrossRef]
17. Prentice, R.L. A log-gamma model and its maximum likelihood estimation. *Biometrika* **1974**, *61*, 539–544. [CrossRef]
18. Lawless, J.F. Inference in the generalized gamma and log gamma distributions. *Technometrics* **1980**, *22*, 409–419. [CrossRef]
19. Young, D.H.; Bakir, S.T. Bias correction for a generalized log-gamma regression model. *Technometrics* **1987**, *29*, 183–191. [CrossRef]
20. Ortega, E.M.M.; Bolfarine, H.; Paula, G.A. Influence diagnostics in generalized log-gamma regression models. *Comput. Stat. Data Anal.* **2003**, *42*, 165–186. [CrossRef]
21. Ortega, E.M.M.; Cancho, V.G.; Paula, G.A. Generalized log-gamma regression models with cure fraction. *Lifetime Data Anal.* **2009**, *15*, 79. [CrossRef] [PubMed]
22. Zhang, D.; Davidian, M. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* **2001**, *57*, 795–802. [CrossRef] [PubMed]
23. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021. Available online: http://www.R-project.org (accessed on 31 July 2021).
24. Tobin, J. Estimation of relationship for limited dependent variables. *Econometrica* **1958**, *26*, 24–36. [CrossRef]
25. Martínez-Flórez, G.; Bolfarine, H.; Gómez, H.W. The alpha–power tobit model. *Commun. Stat.—Theory Methods* **2013**, *42*, 633–643. [CrossRef]
26. Akaike, H. A new look at statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–722. [CrossRef]
27. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
28. Vuong, Q.H. Likelihood ratio tests for models selection and non–nested hypotheses. *Econometrica* **1989**, *57*, 307–333. [CrossRef]
29. Kleiber, C.; Zeileis, A. *Applied Econometrics with R*, 1st ed.; Springer: New York, NY, USA, 2008.
30. Santos, N.; López, R.G.; Israelian, G.; Mayor, M.; Rebolo, R.; García-Gil, A.; De Taoro, M.P.; Randich, S. Beryllium abundances in stars hosting giant planets. *Astron. Astrophys.* **2002**, *386*, 1028–1038. [CrossRef]
31. Tovar-Falón, R.; Bolfarine, H.; Martínez-Flórez G. The Asymmetric Power-Student-t Model for Censored and Truncated Data. *Acad. Bras. Cienc.* **2021**, *93*, e20190920. [CrossRef]
32. Feilgelson, E.D. astrodatR: Astronomical Data. R Package v. 0.1. Available online: https://cran.r-project.org/web/packages/astrodatR/ (accessed on 31 July 2021).