The new odd log-logistic generalized inverse Gaussian regression model

Julio C. S. Vasconcelos [1], **Gauss M. Cordeiro** [2], **Edwin M. M. Ortega** [1], **Elton G. Araújo** [3]

Introdution

The inverse Gaussian (IG) distribution is widely used in several research areas, such as life-time analysis, reliability, meteorology and hydrology, engineering and medicine, among others. Some extensions of the IG distribution have appeared in the literature. For example, the *generalized inverse Gaussian* (GIG) distribution with positive support introduced by Good (1953) in a study of population frequencies. Several papers have investigated the structural properties of the GIG distribution. Sichel (1975) used this distribution to construct mixtures of Poisson distributions. Statistical properties and distributional behavior of the GIG distribution were discussed by Jørgensen (1982) and Atkinson (1985). Dagpunar (1989) provided algorithms for simulating this distribution. Nguyen et al. (2003) showed that it has positive skewness. More recently, Madan et al. (2008) proved that the Black-Scholes formula in finance can be expressed in terms of the GIG distribution function. Koudou and Ley (2014) presented a survey about its characterizations and Lemonte and Cordeiro (2011) obtained some mathematical properties of the *exponentiated generalized inverse Gaussian* (EGIG) distribution. In this paper, we study a new four-parameter model named the *odd log-logistic generalized inverse Gaussian* (OLLGIG) distribution which contains as special cases the GIG and IG distributions, among others. Its major advantage is the flexibility in accommodating several forms of the density function, for instance, bimodal and unimodal shapes. It is also suitable for testing goodness-of-fit of some sub-models. Our main objective is to study a new regression model with two systematic structures based on the OLLGIG distribution e discuss maximum likelihood estimation of the parameters. For these model, we presented some ways to perform global influence (case-deletion) and additionally, we developed residual analysis based on the quantile residual.

1 The OLLGIG distribution

The GIG distribution (JØRGENSEN, 1982) has been applied in several areas of statistical research. The cumulative distribution function (cdf) and probability density function (pdf) of the GIG distribution are given by (for $y > 0$)

$$G_{\mu,\sigma,\nu}(y) = \int_0^y \left(\frac{b}{\mu}\right)^{\nu} \frac{t^{\nu-1}}{2K_{\nu}(\sigma^{-2})} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{b\,t}{\mu} + \frac{\mu}{b\,t}\right)\right] dt \tag{1}$$

and

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

[1]ESALQ, Universidade de São Paulo. e-mail: *juliocezarvasconcelos@hotmail.com*
[2]DEINFO, Universidade Federal de Pernambuco. e-mail: *gausscordeiro@gmail.com*
[1]ESALQ, Universidade de São Paulo. e-mail: *edwin@esalq.usp.br*
[3]UFMS, Universidade Federal de Mato Grosso do Sul. e-mail: *egarauj@yahoo.com.br*

where $\mu > 0$ is the location parameter, $\sigma > 0$ is the scale parameter, $\nu \in \mathbb{R}$ is the shape parameter, $K_\nu(t) = \frac{1}{2}\int_0^\infty y^{\nu-1} \exp\left[-\frac{1}{2}t\,(u+u^{-1})\right] du$ is the modified Bessel function of the third kind and index ν, $b = K_{\nu+1}(\sigma^{-2})/K_\nu(\sigma^{-2})$ and $C = C(\mu,\sigma,\nu) = \left(\frac{b}{\mu}\right)^\nu / 2K_\nu(\sigma^{-2})$.

The statistical literature is filled with hundreds of continuous univariate distributions. Recently, several methods of introducing one or more parameters to generate new distributions have been proposed. Based on the *odd log-logistic generator* (OLL-G) (GLEATON and LYNCH, 2006), we define the OLLGIG cdf, say $F(y) = F(y;\mu,\sigma,\nu,\tau)$, by integrating the log-logistic density function as follows

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

where $\bar{G}_{\mu,\sigma,\nu}(y) = 1 - G_{\mu,\sigma,\nu}(y)$, $\mu > 0$ is a position parameter, $\sigma > 0$ is a scale parameter and $\nu \in \mathbb{R}$ and $\tau > 0$ are shape parameters. Clearly, $G_{\mu,\sigma,\nu}(y)$ is a special case of (3) when $\tau = 1$.

Henceforth, we write $\eta(y) = G_{\mu,\sigma,\nu}(y)$ to simplify the notation. The OLLGIG density function can be expressed as

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

The main motivations for the OLLGIG distribution are to make its skewness and kurtosis more flexible (compared to the GIG model) and also allow bi-modality. We have $\tau = \log\left[\frac{F(y)}{\bar{F}(y)}\right] / \log\left[\frac{\eta(y)}{\bar{\eta}(y)}\right]$, where $\bar{F}(y) = 1 - F(y)$ and $\bar{\eta}(y) = 1 - \eta(y)$.

2 The OLLGIG regression model

In this section, we define the OLLGIG regression model with two systematic structures based on the new distribution. It is a feasible alternative to the GIG and IG regression models for data analysis.

Regression analysis involves specifications of the distribution of Y given a vector $\mathbf{x} = (x_1, ..., x_p)^T$ of covariates. We relate the parameters μ and σ to the covariates by the logarithm link functions

$$\mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}_1) \quad \text{and} \quad \sigma_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}_2), \quad i = 1, \ldots, n, \tag{5}$$

respectively, where $\boldsymbol{\beta}_1 = (\beta_{11}, ..., \beta_{1p})^T$ and $\boldsymbol{\beta}_2 = (\beta_{21}, ..., \beta_{2p})^T$ denote the vectors of regression coefficients and $\mathbf{x}_i^T = (x_{i1}, ..., x_{ip})$. The most important of the parametric regression models defines the covariates in $\mathbf{x}$ which model both μ and σ.

Consider a sample $(y_1, \mathbf{x}_1), ..., (y_n, \mathbf{x}_n)$ of n independent observations. Conventional likelihood estimation techniques can be applied here. The total log-likelihood function

for the vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \nu, \tau)^T$ from model (5) is given by

$$\begin{aligned} l(\boldsymbol{\theta}) &= n\log(\tau) + \nu\sum_{i=1}^{n}\log\left(\frac{b}{\mu_i}\right) + (\nu-1)\sum_{i=1}^{n}\log(y_i) - \sum_{i=1}^{n}\log\left[2K_\nu\left(\frac{1}{\sigma_i^2}\right)\right] - \\ &\quad \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left(\frac{b\,y_i}{\mu_i} + \frac{\mu_i}{b\,y_i}\right) + (\tau-1)\sum_{i=1}^{n}\log\{\eta(y_i)[1-\eta(y_i)]\} - \\ &\quad -2\sum_{i=1}^{n}\log\{\eta(y_i)^\tau + [1-\eta(y_i)]^\tau\}, \end{aligned} \tag{6}$$

where $K_\nu(\cdot)$ and $\eta(\cdot)$ are defined in Section 1. The MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be calculated by maximizing the log-likelihood (6) numerically in the GAMLSS package (STASINOPOULOS et al. 2007) of the **R** software.

We can use the likelihood ratio (LR) statistic for comparing some special sub-models with the OLLGIG regression model. We consider the partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, where $\boldsymbol{\theta}_1$ is a subset of parameters of interest and $\boldsymbol{\theta}_2$ is a subset of remaining parameters. The LR statistic for testing the null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^{(0)}$ versus the alternative hypothesis $H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_1^{(0)}$ is given by $w = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widetilde{\boldsymbol{\theta}})\}$, where $\widetilde{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}$ are the estimates under the null and alternative hypotheses, respectively. The statistic w is asymptotically (as $n \to \infty$) distributed as χ_k^2, where k is the dimension of the subset of parameters $\boldsymbol{\theta}_1$ of interest. For example, the test of $H_0 : \tau = 1$ versus $H : \tau \neq 1$ is equivalent to compare the OLLGIG regression model with the GIG regression model and the LR statistic reduces to $w = 2\left\{l\left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \hat{\nu}, \hat{\tau}\right) - l\left(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \tilde{\nu}, 1\right)\right\}$ where $\widehat{\boldsymbol{\beta}}_1$, $\widehat{\boldsymbol{\beta}}_2$, $\hat{\nu}$ and $\hat{\tau}$ are the MLEs under H and $\tilde{\boldsymbol{\beta}}_1$, $\tilde{\boldsymbol{\beta}}_2$ and $\tilde{\nu}$ are the estimates under H_0.

3 Checking model: diagnostic and residual analysis

A first tool to perform sensitivity analysis, as stated before, is by means of global influence starting from case-deletion (COOK, 1977, 1982). Case-deletion is a common approach to study the effect of dropping the ith observation from the data set. The case-deletion model with systematic structures (5) is given by

$$\mu_l = \exp(\mathbf{x}_l^T\boldsymbol{\beta}_1) \quad \text{and} \quad \sigma_l = \exp(\mathbf{x}_l^T\boldsymbol{\beta}_2), \quad l = 1, 2, \ldots, n, \quad l \neq i. \tag{7}$$

In the following, a quantity with subscript "(i)" means the original quantity with the ith observation deleted. For model (7), the log-likelihood function of $\boldsymbol{\theta}$ is denoted by $l_{(i)}(\boldsymbol{\theta})$. Let $\widehat{\boldsymbol{\theta}}_{(i)} = (\hat{\boldsymbol{\beta}}_{1(i)}^T, \hat{\boldsymbol{\beta}}_{2(i)}^T, \hat{\nu}_{(i)}, \hat{\tau}_{(i)})^T$ be the MLE of $\boldsymbol{\theta}$ from $l_{(i)}(\boldsymbol{\theta})$. To assess the influence of the ith observation on the MLEs $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T, \hat{\nu}, \hat{\tau})^T$, we can compare the difference between $\hat{\boldsymbol{\theta}}_{(i)}$ and $\hat{\boldsymbol{\theta}}$. If deletion of an observation seriously influences the estimates, more attention should be paid to that observation. Hence, if $\hat{\boldsymbol{\theta}}_{(i)}$ is far from $\hat{\boldsymbol{\theta}}$, then the ith observation can be regarded as influential. A first measure of the global influence is defined as the standardized norm of $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$ (generalized Cook distance) given by

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

Another alternative is to assess the values of $GD_i(\boldsymbol{\beta}_1)$, $GD_i(\boldsymbol{\beta}_2)$ and $GD_i(\nu, \tau)$ since these values reveal the impact of the ith observation on the estimates of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and

(ν, τ), respectively. Another popular measure of the difference between $\hat{\boldsymbol{\theta}}_{(i)}$ and $\hat{\boldsymbol{\theta}}$ is the likelihood distance given by

$$LD_i(\boldsymbol{\theta}) = 2\Big\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)})\Big\}.$$

4 Applications

In this section, we provide one application to real data to prove empirically the flexibility of the OLLGIG model. The calculations are performed with the **R** software.

Here, we provide a application of the OLLGIG regression model to evaluation the price of urban residential properties for sale in the municipality of Paranaíba in the State of Mato Grosso do Sul (MS) in Brazil. These data collected in 2017 refer to $n = 45$ houses for sale in the municipality. We construct a OLLGIG regression model with two systematic components to describe the relationship between real estate prices and other explanatory variables, thus allowing an understanding of the behavior of the price variable (BERTRAND, 2002; ARAÚJO et al., 2012). The following response variables and explanatory variables are considered:

- price of the property y_i; this variable was divided by $10,000$;
- area x_{i1} of land in square meters;
- number of parking spaces x_{i2} in the residence (0=no vacancy, 1=one vacancy, 2=more than one vacancy); in this case, two dummy variables, x_{i21} and x_{i22}, are created;
- number of rooms with suites x_{i3} in the residence (0=no suites, 1=one suites, 2=more than one suites); in this case two dummy variables, x_{i31} and x_{i32}, are created;
- if the residence has a swimming pool x_{i4} (0=no, 1=yes);
- if the residence is located in the center of the city x_{i5} (0=no, 1=yes); $\quad i = 1, \ldots, 45$.

We define the OLLGIG regression model by two systematic structures for μ and σ

$$\mu_i = \exp(\beta_{10} + \beta_{11}x_{i1} + \beta_{121}x_{i21} + \beta_{122}x_{i22} + \beta_{131}x_{i31} + \beta_{132}x_{i32} + \beta_{14}x_{i4} + \beta_{15}x_{i5})$$

and

$$\sigma_i = \exp(\beta_{20}+\beta_{21}x_{i1}+\beta_{221}x_{i21}+\beta_{222}x_{i22}+\beta_{231}x_{i31}+\beta_{232}x_{i32}+\beta_{24}x_{i4}+\beta_{25}x_{i5}), \quad i = 1, \ldots, 45.$$

We now consider the test of homogeneity of the scale parameter for the price of urban property data. The LR statistic (see Section 2) for testing the null hypothesis $H_0 : \beta_{21} = \beta_{221} = \beta_{222} = \beta_{231} = \beta_{232} = \beta_{24} = \beta_{25} = 0$ is $w = 31.98$ (p -value < 0.0001), which gives a favorable indication toward to the dispersion not be constant.

The AIC, BIC and global deviance (GD) statistics are listed in Table 1. We note that the OLLGIG regression model presents the lowest AIC, BIC and GD values among the other fitted models. So, there are indications that the OLLGIG model provides a better fit to these data.

Table 1: Goodness-of-fit measures for the the price of urban property data.

Model	AIC	BIC	GD
OLLGIG	322.0612	354.5811	286.0612
GIG	348.8190	379.5323	314.8190
IG	333.3241	362.2307	301.3241

Source: Own author.

Table 2: LR tests for the the price of urban property data.

Models	Hypotheses	Statistic w	p-value
OLLGIG vs GIG	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	28.7579	<0.001
OLLGIG vs IG	$H_0 : \tau = 1$ and $\nu = -0.5$ vs $H_1 : H_0$ is false	15.2629	<0.001

Source: Own author.

We adopt the LR statistics to compare the fitted models in Table 2. We reject the null hypotheses in the two tests in favor of the wider OLLGIG regression model. Rejection is significant at the 5% level and provides clear evidence of the need of the shape parameter τ when modeling real data.

In Table 3, we present the maximum likelihood estimation (MLEs), standard errors (SEs) and p-values for the OLLGIG regression model fitted. The covariates x_2, x_3 and x_5 are significant at the 5% level in the regression structure for the location parameter μ, whereas the covariates x_1, x_3, x_4 and x_5 are significant (at the same level) for the parameter σ. The figures in this table reveal hat the covariate x_1 is not significant with respect to the parameter μ, but it is significant with respect to the parameter σ. This is due to a strong dispersion in the response variable. The covariate x_2 is also significant for the number of parking spaces in the structure of μ. The covariate x_3 is significant in the location and scale structure, i.e., there is a significant difference between the residence that does not have a suite, has a suite or more. The covariate x_4 is not significant in relation to the location, but it is significant in the structure of σ. There is a significant difference in the residence with or without swimming pool for the dispersion parameter.

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 3. The results of such influence measures index plots are displayed in Figure 1. These plots indicate that the cases ♯7, ♯43 and ♯45 are possible influential observations. In addition, Figure 2(a) provide plots of the qrs for the fitted model, thus showing that all observations are in the interval $(-3, 3)$ and a random behavior of the residuals. Hence, there is no evidence against the current suppositions of the fitted model. In order to detect possible departures from the distribution errors in model, as well as outliers, we present the normal plot for the qrs with a generated envelope in Figure 2(b). This plot reveals that the OLLGIG regression model is very suitable for these data, since there are no observations falling outside the envelope. Also, no observation appears as a possible outlier.

5 Concluding remarks

We present a four-parameter distribution called the *odd log-logistic generalized Gaussian inverse* (OLLGIG) distribution, which includes as special cases the generalized Gaus-

Table 3: MLEs, SEs and p-values for OLLGIG regression model fitted.

Parameter	Estimate	SE	p-Value
$\hat{\beta}_{10}$	7.0690	0.4428	<0.001
$\hat{\beta}_{11}$	-0.0005	0.0002	0.0679
$\hat{\beta}_{121}$	0.8069	0.2689	0.0057
$\hat{\beta}_{122}$	0.8407	0.2677	0.0041
$\hat{\beta}_{131}$	-0.8976	0.1945	<0.001
$\hat{\beta}_{132}$	0.4326	0.1872	0.0287
$\hat{\beta}_{14}$	0.5794	0.6941	0.4111
$\hat{\beta}_{15}$	-0.5323	0.1008	<0.001
$\hat{\beta}_{20}$	2.614	0.5982	<0.001
$\hat{\beta}_{21}$	0.0013	9.961e-05	<0.001
$\hat{\beta}_{221}$	-0.2054	0.1316	0.1303
$\hat{\beta}_{222}$	-0.1741	0.1223	0.1660
$\hat{\beta}_{231}$	0.5139	0.0739	<0.001
$\hat{\beta}_{232}$	0.2585	0.1077	0.0235
$\hat{\beta}_{24}$	-2.135	0.4818	<0.001
$\hat{\beta}_{25}$	0.2575	0.0481	<0.001
$\hat{\nu}$	-0.4942	0.1231	<0.001
$\hat{\tau}$	12.764	2.436	

Source: Own author.

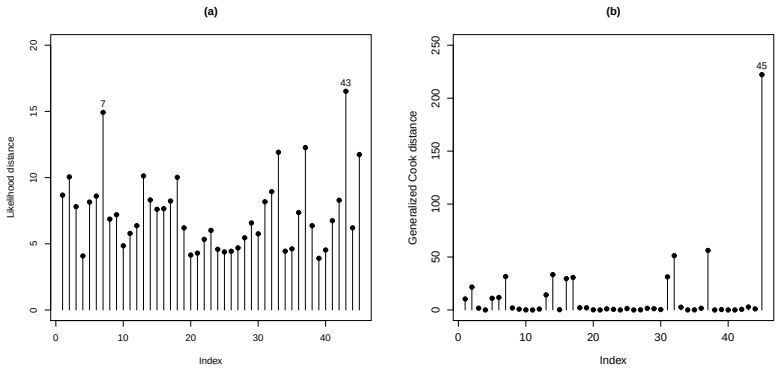

Figure 1: Index plot for $\boldsymbol{\theta}$: (a) $LD_i(\boldsymbol{\theta})$ (likelihood distance) and (b) $GD_i(\boldsymbol{\theta})$ (generalized Cook's distance).

Source: Own author.

sian inverse (GIG) and inverse Gaussian (IG). We provide the OLLGIG regression model with two systematic structures based on this new distribution, which is very suitable for data modeling in which the response variable presents bimodal. The proposed model serves as an important extension to several existing regression models and could be a valuable addition to the literature. The maximum likelihood method is described to estimate the model parameters. Diagnostic analysis is presented to assess overall influences. We also discussed the sensitivity of the maximum likelihood estimates of the adjusted model via quantile residuals. The utility of the proposed OLLGIG regression model is

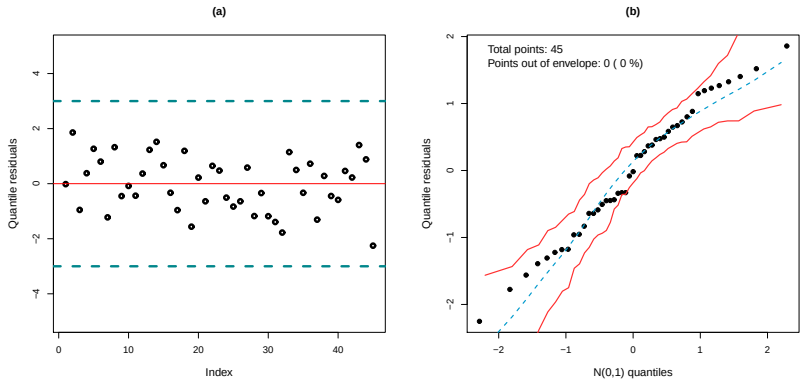

Figure 2: (a) Index plot of the qrs and (b) normal probability plot with envelope for the qrs from the fitted OLLGIG regression model fitted to urban property data.

Source: Own author.

demonstrated by a set of real data for urban residential real estate price data in the municipality of Paranaíba in the state of Mato Grosso do Sul, Brazil.

Acknowledgment

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

References

ATKINSON, A. C. **The simulation of generalized inverse Gaussian and hyperbolic random variables**. SIAM Journal on Scientific and Statistical Computing, v. 3, n. 4, p. 502-515, 1982.

ARAÚJO, E. G. et al. **Proposta de uma metodologia para a avaliação do preço de venda de imóveis residenciais em Bonito/MS baseado em modelos de regressão linear múltipla**. P&D em Engenharia de Produção, v. 10, n. 2, p. 195-207, 2012.

BERTRAND J. W. M.; FRANSOO, Jan C. **Operations management research methodologies using quantitative modeling**. International Journal of Operations and Production Management, v. 22, n. 2, p. 241-264, 2002.

COOK, R. D. **Detection of influential observation in linear regression**. Technometrics, v. 19, n. 1, p. 15-18, 1977.

COOK, R. D.; WEISBERG, S. **Residuals and influence in regression**. New York: Chapman and Hall, 1982.

DAGPUNAR, J. S. **An easily implemented generalised inverse Gaussian generator**. Communications in Statistics-Simulation and Computation, v. 18, n. 2, p. 703-710, 1989.

GLEATON, J. U.; LYNCH, J. D. **Properties of generalized log-logistic families of lifetime distributions**. Journal of Probability and Statistical Science, v. 4, n. 1, p. 51-64, 2006.

GOOD, I. J. **The population frequencies of species and the estimation of population parameters**. Biometrika, v. 40, n. 3-4, p. 237-264, 1953.

JØRGENSEN, B. **Statistical properties of the generalized inverse Gaussian distribution**. Springer-Verlag, New York, 1982.

KOUDOU, A. E.; LEY, C. **Efficiency combined with simplicity: new testing procedures for Generalized Inverse Gaussian models**. Test, v. 23, n. 4, p. 708-724, 2014.

LEMONTE, A. J.; CORDEIRO, G. M. **The exponentiated generalized inverse Gaussian distribution**. Statistics and probability letters, v. 81, n. 4, p. 506-517, 2011.

MADAN, D.; ROYNETTE, B.; YOR, M. **Unifying Black?Scholes Type Formulae Which Involve Brownian Last Passage Times up to a Finite Horizon**. Asia-Pacific Financial Markets, v. 15, n. 2, p. 97-115, 2008.

NGUYEN, T. T. et al. **A proof of the conjecture on positive skewness of generalised inverse Gaussian distributions**. Biometrika, p. 245-250, 2003.

SICHEL, H. S. **On a distribution law for word frequencies**. Journal of the American Statistical Association, v. 70, n. 351a, p. 542-547, 1975.

STASINOPOULOS, D. M. et al. **Generalized additive models for location scale and shape (GAMLSS) in R**. Journal of Statistical Software, v. 23, n. 7, p. 1-46, 2007.