University of São Paulo
"Luiz de Queiroz" College of Agriculture

The parametric, semiparametric and random effect regression model based on the extension of the generalized inverse Gaussian distribution

Julio Cezar Souza Vasconcelos

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2021

Julio Cezar Souza Vasconcelos
Degree in Mathematics

The parametric, semiparametric and random effect regression model based on the extension of the generalized inverse Gaussian distribution

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **EDWIN MOISES MARCOS ORTEGA**

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2021

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Vasconcelos, Julio Cezar Souza

The parametric, semiparametric and random effect regression model based on the extension of the generalized inverse Gaussian distribution/ Julio Cezar Souza Vasconcelos. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2021 .

61 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

. 1. Modelo Aditivo 2. Dado Bimodal 3. Distribuição Gaussiana Inversa Generalizada 4. Gerador Odd Log-Logística 5. Modelo Linear Parcial 6. Regressão com Efeito Aleatório 7. Modelo Semiparamétrico. I. Título

DEDICATION

To my dear parents Maria de Souza Vasconcelos and José Maria Vasconcelos

ACKNOWLEDGMENTS

I thank all those who gave me the opportunity to complete this thesis, especially my parents, Maria de Souza Vasconcelos and José Maria Vasconcelos, my brother Juliano Souza Vasconcelos and my girlfriend Denize Palmito dos Santos, for the love and support.

I also gratefully acknowledge my faculty adviser, Professor Edwin Moises Marcos Ortega, PhD, for the immense help, motivation and enthusiasm during my doctoral program, and the team of the Department of Exact Sciences of ESALQ/USP.

My work was supported financially by the Office to Coordinate Improvement of University Personnel (CAPES) and the National Council for Scientific and Technological Development (CNPq).

SUMMARY

RESUMO

O modelo de regressão paramétrico, semiparamétrico e de efeito aleatório baseado na extensão da distribuição Gaussiana inversa generalizada

Propomos um modelo de regressão baseado na distribuição de quatro parâmetros denominada odd log-logistic Gaussiana inversa generalizada (OLLGIG) com dois componentes sistemáticos adequados para dados unimodais e bimodais que estendam o modelo de regressão GIG heterocedástico. Os modelos de regressão aditivo, parcial ou semiparamétrico podem ser uma opção quando a variável resposta e a variável explicativa tem uma relação não linear, ou seja, não é mais levado em conta uma pressuposição fundamental de linearidade entre essas variáveis. Pensando nisso é proposto três modelos flexíveis denominados de modelos de regressão aditivo, parcial e semiparamétrico baseado na distribuição OLLGIG com uma estrutura sistemática, considerando três diferentes tipos de suavizações penalizadas gerados por splines. Muitos estudos nas áreas de saúde pública, economia, agronomia, medicina, biologia e ciências sociais, entre outros, envolvem observações repetidas de uma variável resposta. A expressão "medidas repetidas" é utilizada para designar medidas obtidas para a mesma variável ou na mesma unidade experimental em mais de uma ocasião. Vários projetos experimentais com medidas repetidas são comuns, como split-plot, crossover e longitudinal. Esses tipos de investigações são denominados estudos de dados correlacionados e desempenham um papel fundamental na análise dos resultados, onde é possível caracterizar alterações nas características de um indivíduo, associando essas variações a um conjunto de covariáveis. Devido à sua natureza, as medidas repetidas possuem uma estrutura de correlação que desempenha um papel importante na análise desses tipos de dados, além disso, a distribuição da variável resposta pode apresentar assimetria ou bimodalidade. Assim, é introduzida uma regressão com intercepto aleatório normal com base na distribuição OLLGIG. Na regressões linear e com efeito aleatório e adotado o método de máxima verossimilhança, já para os modelos: aditivo, parcial e semiparamétrico OLLGIG e utilizado o método de máxima verossimilhança penalizada para estimar os parâmetros do modelos propostos. Além disso, diversas simulações são realizadas para diferentes configurações de parâmetros e tamanhos de amostras para verificar a precisão dos estimadores de máxima verossimilhança e máxima verossimilhança penalizada. São realizadas análises de diagnósticos baseada em case-deletion e resíduos quantílicos. Para comprovar a potencialidade dos modelos de regressão propostos, são realizados ajustes com dados reais.

Palavras-chave: Gerador Odd Log-Logístico, Modelo Aditivo, Modelo Parcial, Modelo Semiparamétrico, Efeito Aleatório

ABSTRACT

The parametric, semiparametric and random effect regression model based on the extension of the generalized inverse Gaussian distribution

We propose a regression model based on the four-parameter distribution called generalized inverse Gaussian odd log-logistic (OLLGIG) with two systematic components suitable for unimodal and bimodal data that extends the heteroscedastic GIG regression model. Additive, partial or semi-parametric regression models can be an option when the response variable and the explanatory variable have a nonlinear relationship, that is, the fundamental assumption of linearity between these variables does not hold. With this in mind, three flexible models are proposed, namely additive, partial and semiparametric regression models based on the OLLGIG distribution with a systematic structure, considering three different types of penalized smoothings generated by splines. Many studies in the areas of public health, economics, agronomics, medicine, biology and social sciences, among others, involve repeated observations of a response variable. The expression " repeated measures " is used to designate measurements obtained for the same variable or in the same experimental unit on more than one occasion. Various experimental designs with repeated measurements exist, such as split-plot, crossover and longitudinal. These types of investigations are called studies of correlated data and play a fundamental role in the analysis of results, where it is possible to characterize changes in the characteristics of an individual by associating these variations to a set of covariates. Due to their nature, the repeated measures have a correlation structure that plays an important role in the analysis of these types of data. In addition, the distribution of the response variable may present asymmetry or bimodality. Thus, a regression with a normal random intercept is introduced based on the OLLGIG distribution. In linear and random regressions, the maximum likelihood method is adopted for the models: additive, partial and semiparametric OLLGIG and the penalized maximum likelihood method are used to estimate the parameters of the proposed models. In addition, several simulations are performed for different parameter configurations and sample sizes to verify the accuracy of the maximum likelihood and penalized maximum likelihood estimators. Diagnostic analyses based on case-deletion and quantile residuals are performed. To prove the potential of the proposed regression models, adjustments are made with real data.

Keywords: Odd Log-Logistic Generator, Additive Model, Partial Model, Semiparametric Model, Random Effect

1 INTRODUTION

The inverse Gaussian (IG) distribution is used to model many phenomena in diverse areas, such as economics, engineering, business, social policy, real estate market, and natural events, among others. An extension of the IG distribution that has been used widely is the generalized inverse Gaussian (GIG), which has positive support. It was initially proposed by Good (1953) in a study of population frequencies. Many papers have examined the structural properties of the GIG distribution. Sichel (1975) used it to produce mixtures of Poisson distributions. The behavior of the GIG distribution and various of its statistical properties were addressed by Jørgensen (1982) and Atkinson (1982). Dagpunar (1989) proposed algorithms to simulate this distribution. Nguyen et al. (2003) pointed out that it has positive asymmetry. Madan et al. (2008) demonstrated that the Black-Scholes formula in finance can be expressed in terms of a function of the GIG distribution. More recently, Koudou and Ley (2014) published a list of its properties and Lemonte and Cordeiro (2011) described some mathematical properties of the exponentiated generalized inverse Gaussian distribution (EGIG).

In the majority of experiments, the response variable is influenced by explanatory variables that elucidate determined characteristics of individuals. Thus, the inclusion of covariables in a regression model is a way to represent the heterogeneity of a population. These covariables, in turn, should be considered in some way in the model to increase its predictive power. The statistical literature contains many types of regression models, such as the semiparametric generalized linear model proposed by Green and Yandell (1985), in which the authors added a nonparametric term in the linear predictor. Another extension of generalized linear models is the generalized additive model (GAM) proposed by Hastie and Tibshirani (1990), in which the term that is controlled in parametric form is altered by an arbitrary function and comes to be controlled in nonparametric form, estimated by smoothed curves (e.g., splines). Rigby and Stasinopoulos (2001) developed generalized additive models for location, scale and shape (GAMLSS), which are widely used in various areas of science. This type of modeling is very flexible, because it allows modeling the location, scale and shape parameters simultaneously.

Several papers have proposed regression models with random effects. Among these works are those of Muniz-Terrera et al. (2016), who developed random effect parametric and nonparametric regressions to analyze cognitive test data; Coupé (2018), who reported advances in statistical modeling in linguistics based in linear mixed-effects regressions; Ho et al. (2019), who presented an analysis of microbiome relative abundance data using a zero-inflated beta GAMLSS model and a meta-analysis of studies using random effects models; Hashimoto et al. (2019), who introduced the random effect log-Burr XII regression; and Dirmeier et al. (2020), who presented host factor prioritization for pan-viral genetic perturbation screens using random intercept models and network propagation. In this sense, in this work we propose parametric, semiparametric and random effect regression models.

For this purpose, our first objective is to define a new four-parameter distribution called the odd log-logistic generalized inverse Gaussian (OLLGIG) to model data pertaining to areas such as the real estate market, engineering and natural phenomena, among others. This model is noteworthy because besides modeling unimodal data, it can also model data where the response variable is bimodal, making it an alternative that in many cases can be more effective than mixture models, in which different situations require two distributions to enable modeling data where the response variable has two modes, making the model more complex. In turn, with respect to semiparametric models, our second objective is to build a regression model based on the OLLGIG distribution that can model unimodal and bimodal data as well as data using extensions of linear regression models, such as the additive, partial and semiparametric cases, in which the different systematic penalized smoothing routines, consisting of splines, are considered in the systematic component. And finally, our third objective is to analyze the correlated data in the presence of bimodality and asymmetry, and based on the studies described, to perform regression with

normal random intercepts based on the OLLGIG for the purpose of considering the possible presence of heterogeneity among some cities in the state of São Paulo, Brazil.

We also describe as special cases the generalized inverse Gaussian (GIG) and inverse Gaussian (IG) distributions, obtain some mathematical properties and discuss the maximum likelihood and penalized maximum likelihood estimation of the parameters. For these models, we present some ways to include global influence (case deletion), and also develop residual analyses based on quantile residuals. Several simulation studies are presented for different configurations of the parameters and sample sizes, and the empirical distribution of the residuals is shown and compared with the standard normal distribution. The results of these studies suggest that the empirical distribution of the quantile residuals for the OLLGIG regression model with two regression structures, along with the additive, partial and semiparametric models, as well as those with random effect on the intercept, have high concordance with the standard normal distribution. This qualifying material is organized as follows.

In Chapter 2, we define a new four-parameter model called the odd log-logistic generalized inverse Gaussian distribution, which extends the generalized inverse Gaussian distributions. We obtain some structural properties of the new distribution and construct an extended regression model based on this distribution with two systematic structures. We adopt the method of maximum likelihood to estimate the model parameters. In addition, various simulations are performed for different parameter settings and sample sizes to check the accuracy of the maximum likelihood estimators. We provide a diagnostics analysis based on case-deletion and quantile residuals. Finally, the potential of the new regression model to predict urban property values is illustrated by means of real data.

In Chapter 3 we propose three flexible regression models, called additive, partial and semiparametric, based on the odd log-logistic generalized inverse Gaussian distribution under three types of penalized smoothing. We adopt the penalized maximum likelihood method to estimate the parameters of the proposed regression models. Furthermore, we present several simulations carried out for different configurations of the parameters and sample sizes to verify the precision of the penalized maximum likelihood estimators. The regression is applied to ethanol data and air quality data.

In Chapter 4, a random effect regression is defined to model correlated data. The maximum likelihood is adopted to estimate the parameters and various simulations are performed for correlated data. Residuals are proposed for the new regression whose empirical distribution is close to normal. The usefulness of the regression is verified based on the average price per hectare of bare land in 10 cities in the state of São Paulo (Brazil).

References

Atkinson, A. (1982). The simulation of generalized inverse gaussian and hyperbolic random variables. *SIAM Journal on Scientific and Statistical Computing*, 3(4):502–515.

Coupé, C. (2018). Modeling linguistic variables with regression models: Addressing non-gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape. *Frontiers in Psychology*, 9:513.

Dagpunar, J. (1989). An easily implemented generalised inverse gaussian generator. *Communications in Statistics-Simulation and Computation*, 18(2):703–710.

Dirmeier, S., Dächert, C., van Hemert, M., Tas, A., Ogando, N. S., van Kuppeveld, F., Bartenschlager, R., Kaderali, L., Binder, M., and Beerenwinkel, N. (2020). Host factor prioritization for pan-viral genetic perturbation screens using random intercept models and network propagation. *PLoS Computational Biology*, 16(2):e1007587.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Green, P. J. and Yandell, B. S. (1985). Semi-parametric generalized linear models. In *Generalized Linear Models*, pages 44–55. Springer.

Hashimoto, E. M., Silva, G. O., Ortega, E. M., and Cordeiro, G. M. (2019). Log-burr xii gamma–weibull regression model with random effects and censored data. *Journal of Statistical Theory and Practice*, 13(2):1–21.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman and Hall.

Ho, N. T., Li, F., Wang, S., and Kuhn, L. (2019). metamicrobiomer: an r package for analysis of microbiome relative abundance data using zero-inflated beta gamlss and meta-analysis across studies using random effects models. *BMC Bioinformatics*, 20(1):188.

Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution*. Springer Science & Business Media, New York.

Koudou, A. E. and Ley, C. (2014). Efficiency combined with simplicity: new testing procedures for generalized inverse gaussian models. *Test*, 23(4):708–724.

Lemonte, A. J. and Cordeiro, G. M. (2011). The exponentiated generalized inverse gaussian distribution. *Statistics & Probability Letters*, 81(4):506–517.

Madan, D., Roynette, B., and Yor, M. (2008). Unifying black–scholes type formulae which involve brownian last passage times up to a finite horizon. *Asia-Pacific Financial Markets*, 15(2):97–115.

Muniz-Terrera, G., Hout, A. v. d., Rigby, R., and Stasinopoulos, D. (2016). Analysing cognitive test data: Distributions and non-parametric random effects. *Statistical Methods in Medical Research*, 25(2):741–753.

Nguyen, T. T., Chen, J. T., Gupta, A. K., and Dinh, K. T. (2003). A proof of the conjecture on positive skewness of generalised inverse gaussian distributions. *Biometrika*, pages 245–250.

Rigby, R. and Stasinopoulos, D. (2001). The gamlss project: a flexible approach to statistical modelling. In *New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling*, volume 337, page 345. University of Southern Denmark.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547.

2 THE NEW ODD LOG-LOGISTIC GENERALIZED INVERSE GAUSSIAN REGRESSION MODEL

Abstract: We define a new four-parameter model called the odd log-logistic generalized inverse Gaussian distribution which extends the generalized inverse Gaussian and inverse Gaussian distributions. We obtain some structural properties of the new distribution. We construct an extended regression model based on this distribution with two systematic structures, which can provide better adjustments to actual data than other special regression models. We adopt the method of maximum likelihood to estimate the model parameters. In addition, various simulations are performed for different parameter settings and sample sizes to check the accuracy of the maximum likelihood estimators. We provide a diagnostics analysis based on case-deletion and quantile residuals. Finally, the potentiality of the new regression model to predict price of urban property is illustrated by means of real data.

Keywords: Generalized inverse Gaussian distribution; Moment; Odd log-logistic generator; Regression model.

2.1 Introduction

The inverse Gaussian (IG) distribution is widely used in several research areas, such as lifetime 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 (1982). 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. We obtain some mathematical properties and 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. For different parameter settings and sample sizes, various simulation studies were performed and the empirical distribution of quantile residual was displayed and compared with the standard normal distribution. These studies suggest that the empirical distribution of the quantile residual for the OLLGIG regression model with two regression structures a high agreement with the standard normal distribution.

This paper is organized as follows. In Section 2.2, we define the OLLGIG distribution. In Section 2.3, we obtain some of its structural properties. We define the OLLGIG regression model in Section 2.4 and evaluate the performance of the maximum likelihood estimators (MLEs) of the model parameters by means of a simulation study. In Section 2.5, we adopt the case-deletion diagnostic measure

and define quantile residuals for the fitted model. Further, we perform various simulations for these residuals. In Section 2.6, we provide two applications to real data to illustrate the flexibility of the OLLGIG regression model. Finally, some concluding remarks are offered in Section 2.7.

2.2 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 $y > 0$)

$$G_{\mu,\sigma,\nu}(y) = \int_0^y \left(\frac{b}{\mu}\right)^{\nu} \frac{t^{\nu-1}}{2K_{\nu}(\sigma^{-2})} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{b\,t}{\mu} + \frac{\mu}{b\,t}\right)\right] dt \tag{2.1}$$

and

$$g_{\mu,\sigma,\nu}(y) = C y^{\nu-1} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{b\,y}{\mu} + \frac{\mu}{b\,y}\right)\right], \tag{2.2}$$

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

We denote by $W \sim \text{GIG}(\mu,\sigma,\nu)$ a random variable having density function (2.2). The mean and variance of W are

$$E(W) = \mu \qquad \text{and} \qquad V(W) = \mu^2\left[\frac{2\,\sigma^2}{b}(\nu+1) + \frac{1}{b^2} - 1\right], \tag{2.3}$$

respectively.

The moment generating function (mgf) of W reduces to

$$M(t) = \left(1 - \frac{2\,\mu\,\sigma^2\,t}{b}\right)^{-\nu/2} K_{\nu}(\sigma^{-2})^{-1} K_{\nu}\left[\frac{1}{\sigma^2}\left(1\frac{2\,\mu\,\sigma^2\,t}{b}\right)^{1/2}\right]. \tag{2.4}$$

We use the re-parameterized GIG distribution according to GAMLSS package in **R** software. For example, we have $GIG(\mu, \sigma\mu^{1/2}, -0.5) = IG(\mu,\sigma)$. Other properties of the GIG distribution are investigated by Jørgensen (1982).

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

$$F(y) = \int_0^{\frac{G_{\mu,\sigma,\nu}(y)}{\bar{G}_{\mu,\sigma,\nu}(y)}} \frac{\tau\,x^{\tau-1}}{(1+x^{\tau})^2} dx = \frac{G_{\mu,\sigma,\nu}(y)^{\tau}}{G_{\mu,\sigma,\nu}(y)^{\tau} + \bar{G}_{\mu,\sigma,\nu}(y)^{\tau}}, \tag{2.5}$$

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 (2.5) 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

$$\begin{aligned} f(y) \;=\; & f(y;\mu,\sigma,\nu,\tau) = \left(\frac{b}{\mu}\right)^{\nu} \frac{\tau\,y^{\nu-1}}{2K_{\nu}\left(\sigma^{-2}\right)} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{by}{\mu} + \frac{\mu}{b\,y}\right)\right] \times \\ & \{\eta(y)[1-\eta(y)]\}^{\tau-1}\left\{\eta(y)^{\tau} + [1-\eta(y)]^{\tau}\right\}^{-2}. \end{aligned} \tag{2.6}$$

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 $\bar{\eta}(y) = 1 - \eta(y)$. Thus, the parameter τ represents the quotient of the log odds ratio for the new and baseline distributions. Note that the pdf and cdf of the OLLGIG distribution depend on integrals, which are calculated numerically in the same way as those of the Birnbaum-Saunders distribution.

Hereafter, we assume that the random variable Y follows the OLLGIG cdf (2.5) with parameters $(\mu, \sigma, \nu, \tau)^T$, say $Y \sim \text{OLLGIG}(\mu, \sigma, \nu, \tau)$. The OLLGIG distribution contains as special cases the GIG distribution when $\tau = 1$ and the IG distribution when $\tau = 1$, $\sigma = \sigma\mu^{1/2}$ and $\nu = -0.5$.

Some plots of the OLLGIG density for selected parameter values are displayed in Figure 2.1. It is evident that the proposed distribution is much more flexible, especially in relation to bi-modality (for $0 < \tau < 1$), than the GIG and IG distributions.

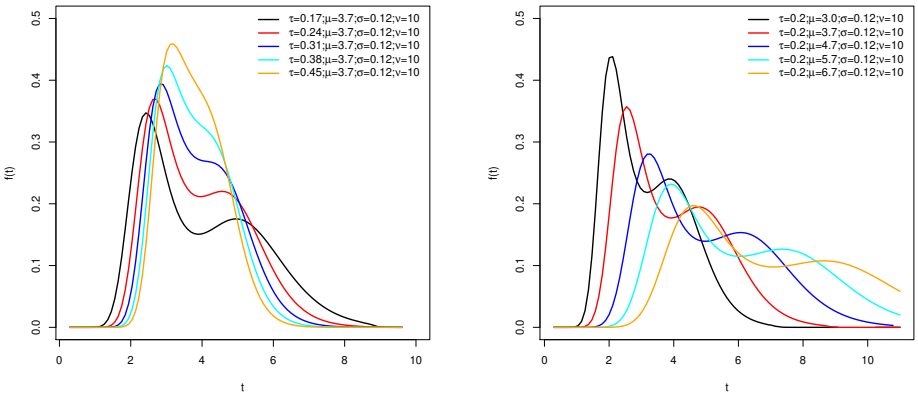

Figure 2.1. Plots of the OLLGIG density for some parameter values.

Equation (2.5) has tractable properties especially for simulations, since its quantile function (qf) takes the simple form

$$y = Q_{GIG}\left(\frac{u^{1/\tau}}{u^{1/\tau} + [1-u]^{1/\tau}}\right), \tag{2.7}$$

where $Q_{GIG}(u) = G^{-1}_{\mu,\sigma,\nu}(u)$ is the qf of the GIG distribution. This scheme is useful because of the existence of fast generators for GIG random variables in some statistical packages. For example, we can fit the generalized additive models for the location, scale, and shape (GAMLSS) (Stasinopoulos et al., 2007) in **R**. We use the GAMLSS package to simulate data from this nonlinear equation. The plots comparing the exact OLLGIG densities and the histograms from two simulated data sets with 100,000 replications for selected parameter values are displayed in Figure 2.2. These plots (and several others not shown here) indicate that the simulated values are consistent with the OLLGIG distribution.

2.3 Properties of the OLLGIG model

2.3.1 Linear representation

By defining the sets $I_i = \{(k, j); k - j = i\}$ for $i = 0, 1, \ldots$, and following the results of Lemonte and Cordeiro (2011), Section 3, we can expand $\eta(y) = G_{\mu,\sigma,\nu}(y)$ as

$$\eta(y) = 1 - \rho - y^{\nu} \sum_{i=0}^{\infty} d_i \, y^i,$$

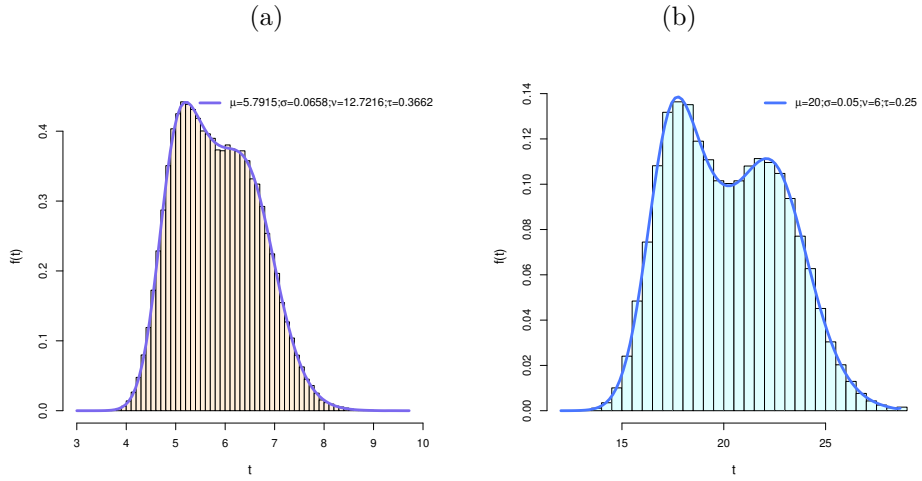

Figure 2.2. Histograms and plots of the OLLGIG density.

where $\rho = \rho(\mu, \sigma, \nu) = C \left(\frac{b}{2\mu\sigma^2}\right)^{-\nu} \sum_{j=0}^{\infty} \Gamma(\nu - j)[-(4\sigma^2)^{-1}]^j / j!$, $d_i = \sum_{(k,j)\in I_i} a_{j,k}$ and

$$a_{j,k} = a_{j,k}(\mu, \sigma, \nu) = \frac{(-1)^{k+j+1}\, C}{(k - j + \nu)\, j!\, k!} \frac{b^{k-j}\, \mu^{j-k}}{2^{k+j}\sigma^{2(k+j)}}.$$

To calculate ρ, the index j can stop after a large number of summands.

Further, we can rewrite $\eta(y)$ after some algebra as

$$\eta(y) = 1 - \rho - c_0 - \sum_{i=1}^{\infty} c_i\, y^i, \tag{2.8}$$

where $c_i = \sum_{k=0}^{i} f_k\, d_{i-k}$ (for $i = 0, 1, \ldots$), $(\nu)_r = \nu(\nu - 1) \cdots (\nu - r + 1)$ is the descending factorial and $f_j = \sum_{r=j}^{\infty} (-1)^{r-j} \binom{r}{j} (\nu)_r / r!$.

We obtain an expansion for $F(y)$ in (2.5). First, we use a power series for $\eta(y)^\tau$ (τ real)

$$\eta(y)^\tau = \sum_{k=0}^{\infty} p_k\, \eta(y)^k, \tag{2.9}$$

where

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

For any real τ, we consider the generalized binomial expansion

$$[1 - \eta(y)]^\tau = \sum_{k=0}^{\infty} (-1)^k \binom{\alpha}{k} \eta(y)^k. \tag{2.10}$$

Inserting (2.9) and (2.10) in Equation (2.5) gives

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

where $q_k = q_k(\tau) = p_k(\tau) + (-1)^k \binom{\tau}{k}$ (for $k \geq 0$). The ratio of the two power series in the last equation can be reduced to

$$F(y) = \sum_{k=0}^{\infty} w_k\, \eta(y)^k, \tag{2.11}$$

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

$$w_k = w_k(\tau) = q_0^{-1} \left(p_k - \sum_{r=1}^{k} q_r \, w_{k-r} \right).$$

By differentiating (2.11), the pdf $f(y)$ reduces to

$$f(y) = \sum_{k=0}^{\infty} w_{k+1} \, h_{k+1}(y), \tag{2.12}$$

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

We can derive a linear representation for $f(y)$ in terms of GIG densities based on the previous results and following the expansions of Lemonte and Cordeiro (2011) that lead to their Equation 2.16. First, we can express $h_{k+1}(y)$ as

$$h_{k+1}(y) = \sum_{j=0}^{k} m_j^{(k)} \, \pi_j(y). \tag{2.13}$$

Here, $\pi_j(y)$ represents the GIG$(\mu, \sigma, \nu + j)$ density function and the coefficients are given by $m_j^{(k)} = (k+1)\, v_{j,k} \, C(\mu, \sigma, \nu)/C(\mu, \sigma, \nu + j)$, where $v_{j,k} = \sum_{i=0}^{k} (-1)^i \binom{k}{i} \sum_{r=0}^{i} \binom{i}{r} \rho^{i-r} \, t_{j,r}$ and the quantities $t_{j,r}$ are determined from the recurrence relation $t_{j,r} = j^{-1} \sum_{m=1}^{j} [(r+1)m - j] \, c_m \, t_{j-m,r}$ (for $j \geq 1$) and $t_{0,r} = 1$ with the c_m's given in Equation (2.8).

By combining (2.12) and (2.13) and changing $\sum_{k=0}^{\infty} \sum_{j=0}^{k}$ by $\sum_{j=0}^{\infty} \sum_{k=j}^{\infty}$, we obtain

$$f(y) = \sum_{j=0}^{\infty} s_j \, \pi_j(y), \tag{2.14}$$

where $s_j = \sum_{k=j}^{\infty} w_{k+1} \, m_j^{(k)}$.

Equation (2.14) reveals that the OLLGIG density function is an infinite linear combination of GIG densities.

2.3.2 Two properties

Equation (2.14) becomes useful in deriving several mathematical properties of the proposed distribution using well-known properties of the GIG distribution. We provide only two examples. The rth moment about zero of the GIG(μ, σ, ν) random variable defined by (2.2) is

$$E(W^r) = \left(\frac{\mu}{b}\right)^r \frac{K_{\nu+r}(\sigma^{-2})}{K_{\nu}(\sigma^{-2})}.$$

Then, the ordinary moments of the OLLGIG random variable Y follow from (2.14) as

$$E(Y^r) = \frac{\mu^r}{K_{\nu}(\sigma^{-2})} \sum_{j=0}^{\infty} s_j \, \frac{K_{\nu+j+r}(\sigma^{-2})}{b_j^r},$$

where $b_j = K_{\nu+j+1}(\sigma^{-2})/K_{\nu}(\sigma^{-2})$.

By combining (2.14) and (2.4), the generating function of Y takes the form

$$M_Y(t) = \frac{1}{K_{\nu}(\sigma^{-2})} \sum_{j=0}^{\infty} s_j \left(1 - \frac{2\,\mu\,\sigma^2\,t}{b_j}\right)^{-\nu/2} K_{\nu+j}\left[\frac{1}{\sigma^2}\left(1 - \frac{2\,\mu\,\sigma^2\,t}{b_j}\right)^{1/2}\right].$$

2.4 The OLLGIG regression model

In many practical applications, the lifetimes are affected by explanatory variables such as sex, smoking, diet, blood pressure, cholesterol level and several others. So, it is important to explore the relationship between the response variable and the explanatory variables. Regression models can be proposed in different forms in statistical analysis. 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

$$\mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}_1) \quad \text{and} \quad \sigma_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}_2), \quad i = 1, \ldots, n, \tag{2.15}$$

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 (2.15) is given by

$$\begin{aligned} l(\boldsymbol{\theta}) \;=\; & n\log(\tau) + \nu\sum_{i=1}^{n}\log\left(\frac{b}{\mu_i}\right) + (\nu-1)\sum_{i=1}^{n}\log(y_i) - \sum_{i=1}^{n}\log\left[2K_\nu\left(\frac{1}{\sigma_i^2}\right)\right] - \\ & \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left(\frac{b\,y_i}{\mu_i} + \frac{\mu_i}{b\,y_i}\right) + (\tau-1)\sum_{i=1}^{n}\log\{\eta(y_i)[1-\eta(y_i)]\} - \\ & 2\sum_{i=1}^{n}\log\{\eta(y_i)^\tau + [1-\eta(y_i)]^\tau\}, \end{aligned} \tag{2.16}$$

where $K_\nu(\cdot)$ and $\eta(\cdot)$ are defined in Section 2.2. The MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be calculated by maximizing the log-likelihood (2.16) numerically in the GAMLSS package of the **R** software. The advantage of this package is that we can adopt many maximization methods, which will depend only on the current fitted model. Initial values for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are taken from the fit of the GIG regression model with $\tau = 1$. We do not have problems of maximizing this log-likelihood function. This fact is shown in Section 4.1, where some simulations of the proposed regression model are given under different scenarios.

Under general regularity conditions, the asymptotic distribution of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is multivariate normal $N_{2p+2}(0, K(\boldsymbol{\theta})^{-1})$, where $K(\boldsymbol{\theta})$ is the expected information matrix. The asymptotic covariance matrix $K(\boldsymbol{\theta})^{-1}$ of $\widehat{\boldsymbol{\theta}}$ can be approximated by the inverse of the $(2p+2) \times (2p+2)$ observed information matrix $-\ddot{\mathbf{L}}(\boldsymbol{\theta})$. The elements of this matrix are calculated numerically. The approximate multivariate normal distribution $N_{2p+2}(0, -\ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}})^{-1})$ for $\widehat{\boldsymbol{\theta}}$ can be used in the classical way to construct approximate confidence for the parameters in $\boldsymbol{\theta}$.

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.

2.4.1 Simulation study

In this part of simulation, we approach in two different ways. First, we perform a simulation to study the behavior of the MLEs of the parameters of the OLLGIG distribution without systematic structures. Second, we evaluate the behavior of the parameter estimates considering two systematic structures.

The OLLGIG distribution Some properties of the MLEs are evaluated using a classical analysis by means of a simulation study. We simulate the OLLGIG distribution as follows:

- Compute the inverse function $F^{-1}(\cdot)$ from the cumulative distribution (2.1).
- Generate $u \sim U(0,1)$.
- Apply u in $F^{-1}(u) = Q(u)$ from Equation (2.7).
- The values $t = Q(u)$ are generated from the OLLGIG distribution, where $Q(u)$ is the inverse of (2.1).

We take $n = 20, 50, 150$ and 350 for each replication, and then evaluate the estimates $\hat{\mu}$, $\hat{\sigma}$, $\hat{\nu}$ and $\hat{\tau}$. We repeat this process $1,000$ times and then calculate the average estimates (AEs), biases and means squared errors (MSEs). In the first scenario, we take $\tau = 0.3662$, $\mu = 5.7915$, $\sigma = 0.0658$ and $\nu = 12.7216$. We use the values fitted in the adjustment to the iris data set in Section 2.6 as shown in the Table 2.4. The estimates of the model parameters are computed using the GAMLSS package in **R** software. The results of the Monte Carlo study under maximum likelihood are given in Table 2.1. They indicate that the MLEs are accurate. Further, the MSEs of the MLEs of the model parameters decay toward zero when n increases in agreement with first-order asymptotic theory.

Table 2.1. AEs, biases and MSEs for the parameters of the OLLGIG distribution.

	scenario 1				scenario 2		
	$n = 20$				$n = 50$		
Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
$\hat{\mu}$	5.9023	0.1109	0.0895	$\hat{\mu}$	5.8618	0.0704	0.0231
$\hat{\sigma}$	0.7468	0.6810	2.6764	$\hat{\sigma}$	0.2119	0.1461	0.5443
$\hat{\nu}$	13.4759	0.7544	66.2484	$\hat{\nu}$	12.6530	-0.0685	7.2487
$\hat{\tau}$	1.0945	0.7283	2.2952	$\hat{\tau}$	0.7115	0.3452	0.4276
	scenario 3				scenario 4		
	$n = 150$				$n = 350$		
Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
$\hat{\mu}$	5.8354	0.0439	0.0039	$\hat{\mu}$	5.8195	0.0281	0.0014
$\hat{\sigma}$	0.0822	0.0165	0.0004	$\hat{\sigma}$	0.0757	0.0099	0.0001
$\hat{\nu}$	12.7241	0.0026	0.0097	$\hat{\nu}$	12.7131	-0.0085	0.0080
$\hat{\tau}$	0.4822	0.1160	0.0242	$\hat{\tau}$	0.4363	0.0700	0.0075

The OLLGIG regression model

We examine the performance of the MLEs in the OLLGIG regression model by means of some simulations with sample sizes $n = 100, 300$ and 500. We simulate $1,000$ samples from two scenarios ($\tau = 0.5$ and $\tau = 1.5$) by considering $\mu_i = \beta_{10} + \beta_{11} x_i$ and $\sigma_i = \beta_{20} + \beta_{21} x_i$. For both cases, we take $\nu = 0.53$. The explanatory variable is generated by $x_i \sim U(0,1)$ and the response variable is generated by $y_i \sim \text{OLLGIG}(\mu_i, \sigma_i, \nu, \tau)$. For each fitted model, we compute the AEs, biases and MSEs. Based on the results given in Table 2.2, we note that the MSEs of the MLEs of β_{10}, β_{11}, β_{20}, β_{21} and τ decay toward zero when the sample size n increases, as usually expected under first-order asymptotic theory. Further, the AEs of the parameters tend to be closer to the true parameter values when n increases. These facts support that the asymptotic normal distribution provides an adequate approximation to the finite sample distribution of the estimates.

Table 2.2. AEs, biases and MSEs for the OLLGIG regression model under scenarios 1 and 2.

scenario 1									
	$n = 100$			$n = 300$			$n = 500$		
Parameter	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
β_{10}	1.5044	0.0044	0.0028	1.5011	0.0011	0.0008	1.5014	0.0014	0.0005
β_{11}	-0.6979	0.0021	0.0100	-0.6981	0.0019	0.0027	-0.6992	0.0008	0.0017
β_{20}	-1.9426	0.0574	0.1223	-1.9606	0.0394	0.0401	-1.9637	0.0363	0.0280
β_{21}	0.3636	0.0136	0.0508	0.3522	0.0022	0.0162	0.3516	0.0016	0.0093
τ	0.5998	0.0998	0.0763	0.5448	0.0448	0.0228	0.5368	0.0368	0.0149
scenario 2									
	$n = 100$			$n = 300$			$n = 500$		
Parameter	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
β_{10}	1.4975	-0.0025	0.0005	1.4982	-0.0018	0.0002	1.4986	-0.0014	0.0001
β_{11}	-0.7017	-0.0017	0.0018	-0.7024	-0.0024	0.0005	-0.7025	-0.0025	0.0003
β_{20}	-2.2627	-0.2627	0.1839	-2.1766	-0.1766	0.0872	-2.1548	-0.1548	0.0659
β_{21}	0.3502	0.0002	0.0717	0.3432	-0.0068	0.0247	0.3454	-0.0046	0.0148
τ	1.2052	-0.2948	0.3226	1.2720	-0.2280	0.1720	1.2932	-0.2068	0.1389

2.5 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) and Cook and Weisberg (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 (2.15) is given by

$$\mu_l = \exp(\mathbf{x}_l^T\boldsymbol{\beta}_1) \quad \text{and} \quad \sigma_l = \exp(\mathbf{x}_l^T\boldsymbol{\beta}_2), \quad l = 1, 2, \ldots, n, \quad l \neq i. \tag{2.17}$$

In the following, a quantity with subscript "(i)"means the original quantity with the ith observation deleted. For model (2.17), 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

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

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

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

Once the model is chosen and fitted, the analysis of the residuals is an efficient way to check the model adequacy. The residuals also serve to identify the relevance of an additional factor omitted from the model and verify if there are indications of serious deviance from the distribution considered for the random error. Further, since the residuals are used to identify discrepancies between the fitted model and the data set, it is convenient to define residuals that take into account the contribution of each observation to the goodness-of-fit measure.

In summary, the residuals allow measuring the model fit for each observation and enable studying whether the differences between the observed and fitted values are due to chance or to a systematic behavior that can be modeled. The quantile residuals (qrs) (Dunn and Smyth, 1996) for the OLLGIG regression model with two systematic structures are defined by

$$qr_i = \Phi^{-1}\left\{\frac{\eta(y_i)^\tau}{\eta(y_i)^\tau + [1 - \eta(y_i)]^\tau}\right\}, \tag{2.18}$$

where $\eta(\cdot)$ is given in Equation (2.1) and $\Phi(\cdot)^{-1}$ is the inverse cumulative standard normal distribution.

Atkinson (1985) suggested the construction of an envelope to have a better interpretation of the probability normal plot of the residuals. The simulated confidence bands of the envelope should contain the residuals. If the model is well-fitted, the majority of points will be within these bands and randomly distributed. The construction of the confidence bands follows the steps:

- Fit the proposed model and calculate the residuals qr_i's;
- Simulate k samples of the response variable using the fitted model;
- Fit the model to each sample and calculate the residuals qr_{ij} $(j = 1, \ldots, k$ and $i = 1, \ldots, n)$;
- Arrange each group of n residuals in rising order to obtain $qr_{(i)j}$ for $j = 1, \ldots, k$ and $i = 1, \ldots, n$;
- For each i, calculate the mean, minimum and maximum $qr_{(i)j}$, namely

$$qr_{(i)M} = \sum_{j=1}^{k} \frac{qr_{(i)j}}{k}, \quad qr_{(i)I} = \min\{qr_{(i)j} : 1 \leq j \leq k\} \text{ and } qr(i)S = \max\{qr_{(i)j} : 1 \leq j \leq k\};$$

- Include the means, minimum and maximum together with the values of qr_i against the expected percentiles of the standard normal distribution.

The minimum and maximum values of the $qr_i's$ form the envelope. If the model under study is correct, the observed values should be inside the bands and distributed randomly.

Simulation study

A simulation study is conducted to investigate the behavior of the empirical distribution of the qrs for the OLLGIG regression model. We generate $1,000$ samples based on the algorithm presented in Section 2.4.1. We also give the normal probability plots to assess the degree of deviation from the normality assumption of the residuals. Based on the plots in Figures 2.3 and 2.4 representing the first and second scenarios, respectively, we conclude that the empirical distribution of the qrs agrees with the standard normal distribution in both scenarios. This empirical distribution becomes closer to the standard normal distribution when n increases in both scenarios.

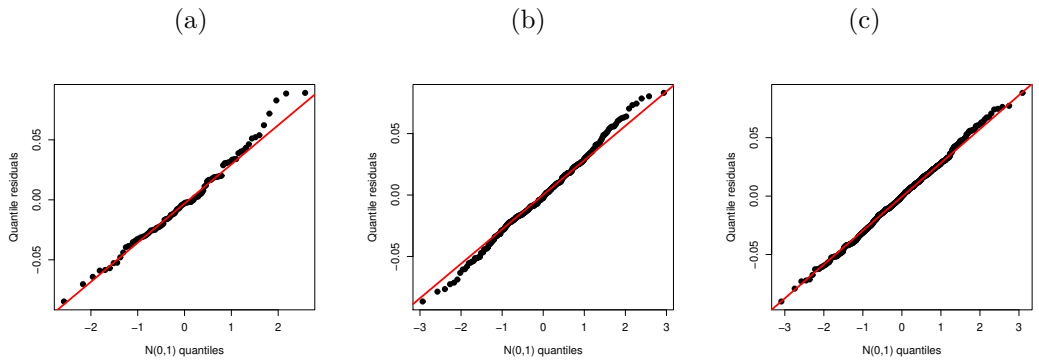

Figure 2.3. Normal probability plots for qr_i in the OLLGIG regression model under scenario 1 ($\tau = 0.5$) (a) $n = 100$. (b) $n = 300$. (c) $n = 500$.

2.6 Applications

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

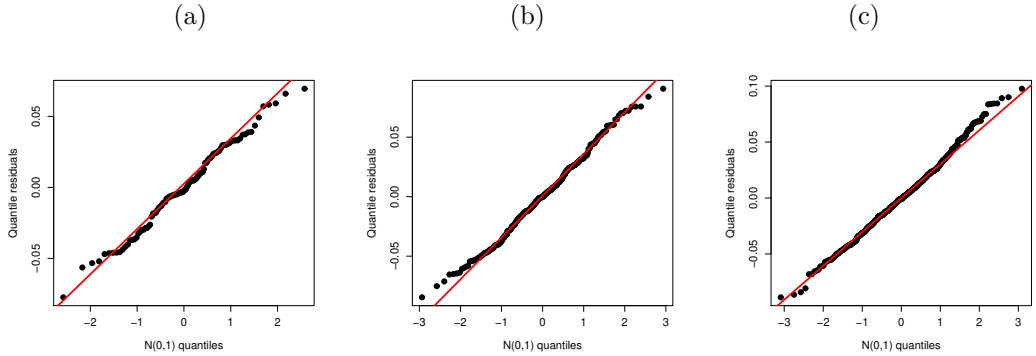

Figure 2.4. Normal probability plots for qr_i in the OLLGIG regression model for scenario 2 ($\tau = 1.5$) (a) $n = 100$. (b) $n = 300$. (c) $n = 500$.

2.6.1 Application 1: Iris data

In the first application, the OLLGIG distribution is compared with the nested GIG and IG distributions. The data set is iris, in which it provides measurements in centimeters of the variables length and width of the septal and length and width of the petal, respectively, for 50 flowers of each of the 3 iris species (setosa , versicolor and virginica). In this application, the variable septum length (Sepal.Length) is used. This data set has been analyzed by several authors in multivariate analysis, for example, Anderson (1935) and Fisher (1936). We show that the distribution for these data presents bi-modality.

Table 2.3 provides a descriptive summary for these data and indicate positively distorted distributions with varying degrees of variability, skewness, and kurtosis.

Table 2.3. Descriptive statistics for iris flower data.

Mean	Median	SD	Skewness	Kurtosis	Min.	Max.
5.843	5.800	0.828	0.3086	-0.6058	4.300	7.900

A brief descriptive analysis of the data in Table 2.3 reveals that the average score of the variable septum length is 5.843, the median value is 5.800, thus indicating that the data has a symmetric distribution.

In Table 2.4, we report the MLEs of the model parameters and their standard errors (SEs) in parentheses. We give in Table 2.5 the following goodness-of-fit measures: Akaike Information Criterion (AIC), Consistent Akaike Information Criterion (CAIC), Bayesian Information Criterion (BIC), Hernnan-Quinn Information Criterion (HIQC), Cramér-von Misses (W^*), Anderson Darling (A^*) and Kolmogarov-Smirnov (KS) test statistic. The small values of these measures, the better the fit. The figures in Table 2.5 indicate that the OLLGIG distribution has the lowest values of AIC, CAIC, BIC, HQIC, A^*, W^* and KS among those of the fitted models and therefore it could be chosen as the best model.

Table 2.4. MLEs and SEs (in parentheses) of the model parameters for the iris data.

Model	τ	μ	σ	ν
OLLGIG	0.3662	5.7915	0.0658	12.7216
	(0.0685)	(0.0091)	(0.0079)	(0.0130)
GIG	1	5.8433	0.1413	0.1000
	(-)	(0.0674)	(0.0082)	(72.9562)
IG	1	5.8433	0.0585	-0.5
	(-)	(0.0674)	(0.0034)	(-)

Table 2.5. Goodness-of-fit measures for the iris data.

Model	AIC	CAIC	BIC	HQIC	A^*	W^*	KS
OLLGIG	365.0638	365.3397	377.1064	369.9563	0.3474	0.0486	0.0578
GIG	369.8170	369.9814	378.8489	373.4864	0.7242	0.1164	0.0881
IG	367.8134	367.8951	373.8347	370.2597	0.7244	0.1165	0.0881

We consider LR statistics to compare nested models. The OLLGIG distribution includes some sub-models as mentioned above, thus allowing their evaluations relative to the others and to a more general model. The values of the LR statistics are listed in Table 2.6. It is evident from the figures in this table that the OLLGIG distribution outperforms its sub-models according to the values of the LR statistics. So, it indicates that the OLLGIG model provides a better fit to these data than their sub-models.

Table 2.6. LR tests for the iris data.

Models	Hypotheses	Statistic w	p-value
OLLGIG vs GIG	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	6.7532	0.0094
OLLGIG vs IG	$H_0 : \tau = 1$ and $\nu = -0.5$ vs $H_1 : H_0$ is false	6.7496	0.0342

More information is provided by a visual comparison of the histogram of the data and the fitted density functions and cumulative functions. The plots of the fitted OLLGIG, GIG and IG densities are displayed in Figure 2.5(a). The estimated OLLGIG density provides the closest fit to the histogram of the data. In order to assess if the model is appropriate, the plots of the fitted OLLGIG, GIG and IG cumulative distributions and the empirical cdf are displayed in Figure 2.5(b). They indicate that the OLLGIG distribution provides a good fit to these data.

(a) (b)

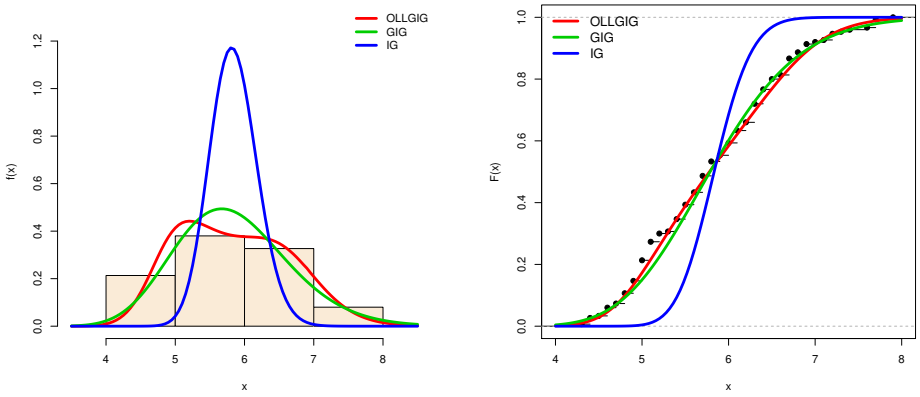

Figure 2.5. (a) Estimated densities of the OLLGIG, GIG and IG models for iris data. (b) Estimated cumulative functions of the OLLGIG, GIG and IG models and the empirical cdf for iris data.

In Figure 2.6, we note that the iris data has a bi-modality shape, where the GIG and IG distributions they can not have (see Figure 2.5(a)).

2.6.2 Application 2: Price of urban property data.

Here, we provide a second 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.

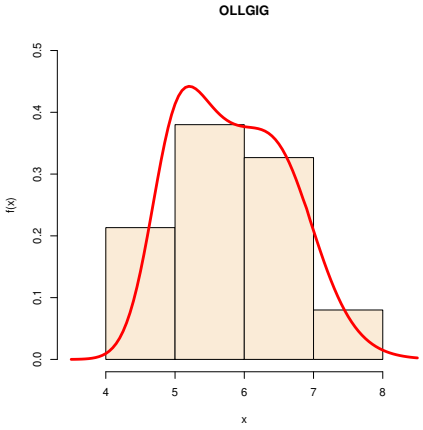

Figure 2.6. Estimated densities of the OLLGIG for iris data.

In the context of real estate appraisal, it is necessary to develop statistical methodologies (characterized by the scientific accuracy) of residential property prices. Besides this aspect, we can perceive the rare use of such methodologies by the real estate market. 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 (Will M. Bertrand and Fransoo, 2002) and (Araújo et al., 2012). The response variable and explanatory variables are considered as follows:

- 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); $\quad i = 1, \ldots, 45$.

In the descriptive analysis of the data from Table 2.7, the mean score of the variable value is 24.98, which is not close to the median value 17.00, thus indicating that the data has an asymmetric distribution.

Table 2.7. Descriptive analysis of the price of urban property data$.

Mean	Median	SD	Skewness	Kurtosis	Min.	Max.
24.98	17.00	23.9180	3.3330	14.0134	5.50	150.00

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 2.4) for testing the null hypothesis $H_0 : \beta_{21} = \beta_{221} = \beta_{222} = \beta_{231} = \beta_{232} = \beta_{24} = \beta_{25} = 0$ is $w = 31.98\,(p \text{ value} < 0.0001)$, which gives a favorable indication toward to the dispersion not be constant.

In Table 2.8, we present the MLEs, SEs and p-values. 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. This fact can also be noted in Figure 2.7(a). The covariate x_5 is significant in relation to both parameters μ and σ, i.e., there is a significant difference between the residence being in the center of the city and outside the center. This fact can also be noted in Figure 2.7(b).

Table 2.8. MLEs, standard errors (SEs) and p-values for the OLLGIG regression model fitted for the price of urban property data.

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	

The AIC, BIC and global deviance (GD) statistics are listed in Table 2.9. 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 2.9. 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

(a) (b)

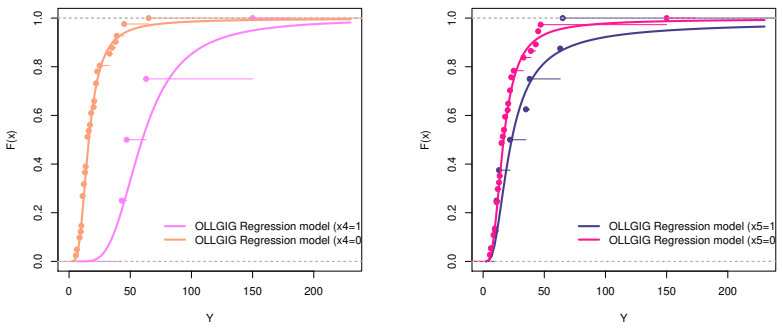

Figure 2.7. Estimated cdf from the fitted OLLGIG regression model and the empirical cdf for the the price of urban property data. (a) For covariate x_4, and (b) For covariate x_5.

We adopt again the LR statistics to compare the fitted models in Table 2.10. 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.

Table 2.10. 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

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 2.5. The results of such influence measures index plots are displayed in Figure 2.8. These plots indicate that the cases ♯7, ♯43 and ♯45 are possible influential observations.

(a) (b)

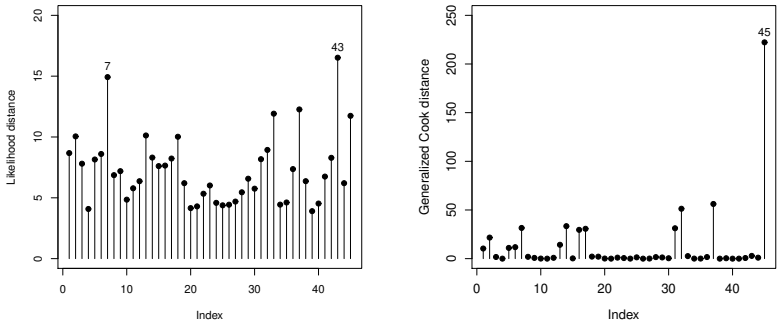

Figure 2.8. Index plot for $\boldsymbol{\theta}$: (a) $LD_i(\boldsymbol{\theta})$ (likelihood distance) and (b) $GD_i(\boldsymbol{\theta})$ (generalized Cook's distance).

In addition, Figure 2.9(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 2.9(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.

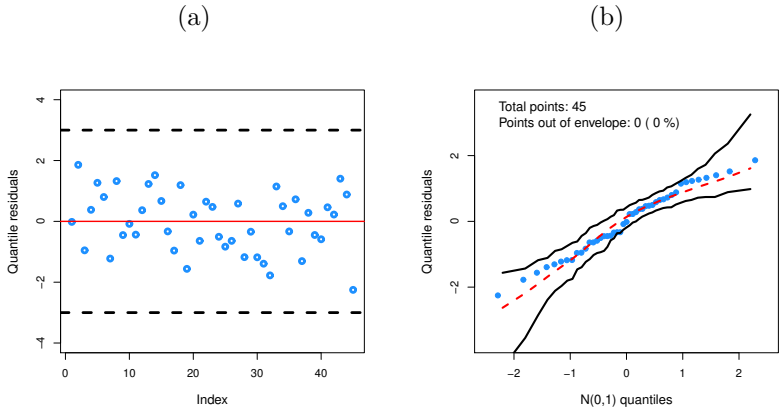

Figure 2.9. (a) Index plot of the qrs and (b) normal probability plot with envelope for the qrs from the fitted OLLGIG regression model fitted to urban property data.

2.7 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 Gaussian inverse (GIG) and inverse Gaussian (IG). We provide some of its mathematical properties. Further, we define the OLLGIG regression model with two systematic structures based on this new distribution, which is very suitable for modeling censored and uncensored data. The proposed model serves as an important extension to several existing regression models and could be a valuable addition to the literature. Some simulation are performed for different parameter settings and sample sizes. The maximum likelihood method is described for estimating the model parameters. Diagnostic analysis is presented to assess global influences. We also discuss the sensitivity of the maximum likelihood estimates from the fitted model via quantile residuals. The utility of the proposed OLLGIG regression model is demonstrated by means of a real data set for price data of urban residential properties in the municipality of Paranaíba in the State of Mato Grosso do Sul, Brazil.

References

Anderson, E. (1935). The irises of the gaspe peninsula. *Bulletin of the American Iris Society*, 59:2–5.

Araújo, E. G., Pereira, J. C., Ximenes, F., Spanhol, C. P., Garson, S., and Araújo, E. (2012). 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*, 10(2):195–207.

Atkinson, A. (1982). The simulation of generalized inverse gaussian and hyperbolic random variables. *SIAM Journal on Scientific and Statistical Computing*, 3(4):502–515.

Atkinson, A. C. (1985). Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis. Technical report.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.

Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Dagpunar, J. (1989). An easily implemented generalised inverse gaussian generator. *Communications in Statistics-Simulation and Computation*, 18(2):703–710.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Gleaton, J. and Lynch, J. (2006). Properties of generalized log-logistic families of lifetime distributions. *Journal of Probability and Statistical Science*, 4(1):51–64.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution*. Springer Science & Business Media, New York.

Koudou, A. E. and Ley, C. (2014). Efficiency combined with simplicity: new testing procedures for generalized inverse gaussian models. *Test*, 23(4):708–724.

Lemonte, A. J. and Cordeiro, G. M. (2011). The exponentiated generalized inverse gaussian distribution. *Statistics & Probability Letters*, 81(4):506–517.

Madan, D., Roynette, B., and Yor, M. (2008). Unifying black–scholes type formulae which involve brownian last passage times up to a finite horizon. *Asia-Pacific Financial Markets*, 15(2):97–115.

Nguyen, T. T., Chen, J. T., Gupta, A. K., and Dinh, K. T. (2003). A proof of the conjecture on positive skewness of generalised inverse gaussian distributions. *Biometrika*, pages 245–250.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547.

Stasinopoulos, D. M., Rigby, R. A., et al. (2007). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46.

Will M. Bertrand, J. and Fransoo, J. C. (2002). Operations management research methodologies using quantitative modeling. *International Journal of Operations & Production Management*, 22(2):241–264.

3 THE SEMIPARAMETRIC REGRESSION MODEL FOR BIMODAL DATA WITH DIFFERENT PENALIZED SMOOTHERS APPLIED TO CLIMATOLOGY, ETHANOL AND AIR QUALITY DATA

Abstract: Semiparametric regressions can be used to model data when covariables and the response variable have a nonlinear relationship. In this work, we propose three flexible regression models for bimodal data called the additive, additive partial and semiparametric regressions, basing on the *odd log-logistic generalized inverse Gaussian* distribution under three types of penalized smoothers, where the main idea is not to confront the three forms of smoothings, but to show the versatility of the distribution with three types of penalized smoothers. We present several Monte Carlo simulations carried out for different configurations of the parameters and some sample sizes to verify the precision of the penalized maximum likelihood estimators. The usefulness of the proposed regressions is proved empirically through three applications to climatology, ethanol and air quality data.
Keywords: Additive model; additive partial model; generalized inverse Gaussian distribution; semiparametric model; splines.

3.1 Introduction

For many years, the normal linear regression model has been used to explain most random phenomena. Even when the phenomenon under study does not present a response for which the normality assumption is reasonable, some types of transformations are suggested to achieve the desired normal distribution. Another important problem in regression models occurs when there are linear and nonlinear effects on the response variable in a single data set.

A great effort was undertaken to provide more flexible assumptions so that these regressions could model real situations with greater precision. However, these flexible assumptions lead to more complex regression models which are very hard to be interpreted in some cases. Nowadays, the literature has various types of regression models such as the generalized linear semiparametric models pioneered by Green and Yandell (1985), where it was added a nonparametric term to the linear predictor. Another extension of the generalized linear models is the generalized additive model (GAM) introduced by Hastie and Tibshirani (1990), in which the term that is controlled in parametric form is altered by an arbitrary function and becomes controlled in nonparametric form, and then it is estimated by smoothed curves (such as splines). Ruppert et al. (2003) demonstrate that nonparametric regression can be considered as a relatively simple extension of parametric regression and combine the two together, in what refers to semiparametric regression, they approach semiparametric regression based on penalized regression splines and mixed models. Rigby and Stasinopoulos (2005) developed a generalized additive model for location, scale and shape (GAMLSS), which has been widely used in various areas of science due to its flexibility, by allowing modeling the location, scale and shape simultaneously. The utility of the semiparametric regression method in scenarios of real change is of extreme importance. For example, Fan and Hyndman (2011) proposed a new statistical method to predict short-term electricity demand based on a semiparametric additive model, Lebotsa et al. (2018) presented an application of partially linear additive quantile regression models to predict short-term electricity demand using data from South Africa, Hudson et al. (2010) showed the benefits of the GAMLSS in the modeling and interpretation of possible nonlinear climate impacts on eucalyptus tree growth, Del Giudice et al. (2015) presented a hedonic price function constructed through a semiparametric additive model, and more recently, Etienne et al. (2019) utilized a semiparametric model and stochastic frontier model to estimate the efficiency of corn production by smallholders in Zimbabwe.

On the other hand, the distributions commonly used in regression models are being modified

and/or generalized to enable them to model different complex forms of data. Hence, it is convenient to consider parametric families of distributions that are flexible enough to capture a wide range of symmetric, asymmetric and bimodal behaviors.

In this article, we adopt as baseline the *odd log-logistic generalized inverse Gaussian* (OLLGIG) distribution introduced recently by Souza Vasconcelos et al. (2019). Thus, the fundamental objective is propose additive, additive partial and semiparametric regression models for bimodal data from in the OLLGIG distribution with different penalized smoothers.

The inferential component is carried out using the asymptotic distribution of the maximum likelihood estimators (MLEs). These models are presented with some methods to effect global influence. Additionally, we develop residual analysis from quantile residuals (qrs). For some parameter settings, additive terms and sample sizes, diverse Monte Carlo simulations are carried out making comparisonthe empirical distribution of the qrs with the standard normal distribution. These simulations indicate that the empirical distribution of these residuals with different penalized smoothers present conformity in what it refers to standard normal distribution.

The rest of the paper is structured following way. In Section 3.2, the OLLGIG semiparametric regression model will be defined based on different penalized smoothers, estimate their parameters by the penalized maximum likelihood method, diagnostic and residual analysis are discussed. In Section 3.3 some properties of the maximum likelihood estimators are evaluated using a simulation study. In Section 3.4, we show empirically how flexible, practical relevance and applicability of the presented regression models by means of three real data sets. Section 3.5 is devoted to some concluding remarks.

3.2 The OLLGIG semiparametric regression

For modelling OLLGIG distributions, GAMLSS package (Stasinopoulos et al., 2007) available in **R** software was used, implementing a new distribution, as described in Section 4.2 in Stasinopoulos et al. (2008). For the regression analysis we use the function *gamlss(·)* from the GAMLSS package (Stasinopoulos et al., 2007), in which the regression structures with the penalized smoothers are described in Tables 3.5, 3.10 and 3.13.

The inverse Gaussian (IG) and generalized inverse Gaussian (GIG) distributions are highly applied in various areas of science for example the survival and reliability analysis, meteorology, hydrology and engineering, among others. Recently, Souza Vasconcelos et al. (2019) defined the general form for the OLLGIG cdf, is (for $y > 0$)

$$F(y) = F(y;\mu,\sigma,\nu,\tau) = \frac{G_{\mu,\sigma,\nu}(y)^{\tau}}{G_{\mu,\sigma,\nu}(y)^{\tau} + [1 - G_{\mu,\sigma,\nu}(y)]^{\tau}}, \tag{3.1}$$

where

$$G_{\mu,\sigma,\nu}(y) = \int_0^y \left(\frac{b}{\mu}\right)^{\nu} \frac{t^{\nu-1}}{2K_{\nu}(\sigma^{-2})} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{b\,t}{\mu} + \frac{\mu}{b\,t}\right)\right] dt \tag{3.2}$$

where $G_{\mu,\sigma,\nu}(y)$ is the cdf of the GIG distribution, $\mu > 0$ represents the average of the GIG distribution, $\sigma > 0$ is a scale parameter, $\nu \in \mathbb{R}$ and $\tau > 0$ are shape parameters, $b = K_{\nu+1}(\sigma^{-2})/K_{\nu}(\sigma^{-2})$, and $K_{\nu}(t) = \frac{1}{2}\int_0^{\infty} y^{\nu-1} \exp\left[-\frac{1}{2}t\left(u + u^{-1}\right)\right] du$ is the modified Bessel function of the third kind and index ν. Clearly, $G_{\mu,\sigma,\nu}(y)$ is a special case of (3.1) when $\tau = 1$. Further details and properties of the GIG distribution can be found in Jørgensen (1982).

We write $\eta(y) = G_{\mu,\sigma,\nu}(y)$ to simplify the notation. Then, the OLLGIG density function (for $y > 0$) can be written as

$$\begin{aligned} f(y) \;=\; & f(y;\mu,\sigma,\nu,\tau) = \left(\frac{b}{\mu}\right)^{\nu} \frac{\tau\, y^{\nu-1}}{2K_{\nu}\left(\sigma^{-2}\right)} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{by}{\mu} + \frac{\mu}{b\,y}\right)\right] \\ & \times\{\eta(y)[1-\eta(y)]\}^{\tau-1}\left\{\eta(y)^{\tau} + [1-\eta(y)]^{\tau}\right\}^{-2}. \end{aligned} \tag{3.3}$$

The main properties and motivations of the OLLGIG distribution is that it is more flexible in relation to asymmetry and kurtosis as well as allowing bi-modality when $0 < \tau < 1$. If Y is a random variable with density (3.3), then we write $Y \sim \text{OLLGIG}(\mu, \sigma, \nu, \tau)$. The OLLGIG distribution contains two important special cases, the GIG distribution when $\tau = 1$ and the IG distribution when $\tau = 1$, $\sigma = \sigma\mu^{1/2}$ and $\nu = -0.5$.

The OLLGIG model is easily simulated, since its quantile function (qf) takes the simple form $y = Q_{GIG}\left(\frac{u^{1/\tau}}{u^{1/\tau}+[1-u]^{1/\tau}}\right)$, where $Q_{GIG}(u) = G^{-1}_{\mu,\sigma,\nu}(u)$ is the qf of the GIG distribution.

In many research areas there are continuous explanatory variables with nonlinear effects in the response variable and more flexible models under less restrict assumptions are desirable. So, a non-parametric approach in one or more covariables may be a suitable choice to control the effects of the continuous covariables, or even to explain nonlinear tendencies of these variables. In this context, we propose three semiparametric regressions based on the OLLGIG distribution, namely: the OLLGIG additive regression, the OLLGIG additive partial regression and the OLLGIG semiparametric regression with different penalized smoothers. The likelihood ratio (LR) statistics can be adopted to discriminate among the OLLGIG, GIG and IG semiparametric regressions. The penalized likelihood function is used to fit the OLLGIG semiparametric regression.

Regression analysis involves specifications for the distribution of Y_i given a vector of covariables $\mathbf{w}_i = (w_{i1}, \ldots, w_{ip})^T$. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})^T$ (for $i = 1, \ldots, n$) be the vector of covariables that has a nonlinear form with the response variable. The important of the OLLGIG semiparametric regression defines the parameters depending on $\mathbf{w}_i$ and $\mathbf{x}_i$. The μ_i parameters are related to covariables by the link functions:

$$\mu_i = \exp\left[\sum_{\xi=1}^{J} h_\xi(x_{i\xi})\right] \quad \rightarrow \quad \text{OLLGIG additive regression model;} \tag{3.4}$$

$$\mu_i = \exp\left[\mathbf{w}_i^T\boldsymbol{\beta} + h(x_{i\xi})\right] \quad \rightarrow \quad \text{OLLGIG additive partial regression model;} \tag{3.5}$$

$$\mu_i = \exp\left[\mathbf{w}_i^T\boldsymbol{\beta} + \sum_{\xi=1}^{J} h_\xi(x_{i\xi})\right] \quad \rightarrow \quad \text{OLLGIG semiparametric regression model,} \tag{3.6}$$

where $h_\xi(\cdot)$ is the smooth function related to the continuous explanatory variable with non-linear effects that are non-parametric controlled (for $i = 1, \ldots, n$, $\xi = 1, \ldots, J$) and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the full parameter vector, note that equations (3.4) and (3.5) are particular cases of equation (3.6).

In this article, we shall discuss two smoothing functions called cubic spline and P-spline in the systematic structure.

- **Cubic spline**

  The cubic spline is represented by the cs$(\cdot)$ function, which uses the *smooth.spline(·)* command to smooth a curve available in the GAMLSS package (Stasinopoulos et al., 2007). Let $y_1, \ldots, y_n$ be n observations from the OLLGIG$(\mu_i, \sigma, \nu, \tau)$ distribution. For the semiparametric models (3.4), (3.5) and (3.6), the fixed and random effects $\boldsymbol{\theta}$ and $\mathbf{h}$, respectively, are estimated by maximizing the penalized log-likelihood function (see, for instance, Hastie and Tibshirani (1990) and Green and Silverman (1993)) has the form

$$l_p(\boldsymbol{\theta}, \mathbf{h}) \quad = \quad l(\boldsymbol{\theta}) - \sum_{\xi=1}^{J} \frac{\lambda_\xi}{2}\, \mathbf{h}_\xi^T \mathbf{K}_\xi \mathbf{h}_\xi. \tag{3.7}$$

where

$$\begin{aligned} l(\boldsymbol{\theta}) &= n\log(\tau) + \nu\sum_{i=1}^{n}\log\left(\frac{b}{\mu_i}\right) + (\nu-1)\sum_{i=1}^{n}\log(y_i) - n\log\left[2K_\nu\left(\frac{1}{\sigma^2}\right)\right] \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{b\,y_i}{\mu_i}+\frac{\mu_i}{b\,y_i}\right) + (\tau-1)\sum_{i=1}^{n}\log\{\eta(y_i)[1-\eta(y_i)]\} \\ &\quad -2\sum_{i=1}^{n}\log\{\eta^\tau(y_i)+[1-\eta(y_i)]^\tau\}, \end{aligned} \tag{3.8}$$

$\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma, \nu, \tau)^T$ is the parameter vector, $\lambda_\xi > 0$ is the smoothing parameter, which characterizes the smoothness of the curve, i.e., it controls the quality of the curve fitting, for the vector of smoothed function $\mathbf{h}_\xi = (h_\xi(x_{1\xi}), \ldots, h_\xi(x_{q\xi}))^T$, where q are the distinct and ordered observations of the covariable that is controlled in a non-parametric way, with $\xi = 1, \ldots, J$ number of covariables of non-linear effect on $y_i (i = 1, \ldots, n)$, $\mathbf{K}_\xi = \mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^T$ is a $q \times q$ definite positive matrix, where $\mathbf{Q}$ is a matrix of order $q \times (q-2)$ and $\mathbf{R}$ is a matrix of order $(q-2) \times (q-2)$. For more details, see Green and Silverman (1993).

Equation (3.7) is a general form of writing the penalized log-likelihood function of the semiparametric regression models.

- If we do not consider the systematic form $\mathbf{w}_i^T\boldsymbol{\beta}$, then Equation (3.7) refers to the penalized log-likelihood function for the OLLGIG additive regression,
- If $\xi = 1$, it is the penalized log-likelihood function for the OLLGIG additive partial regression,
- If $\xi > 1$, it refers to the log-likelihood function for the OLLGIG semiparametric regression.

- **P-spline**

The other smoothing function used in the paper is called the P-spline (Eilers and Marx, 1996), which involves penalized splines, more particularly the ps(·) and pb(·) functions. The smoothed ps(·) function is based on the function of Brian Marx, while the smoothed pb(·) function follows the function defined by Paul Eilers. We present two main differences the ps(·) and pb(·) functions:

- the ps(·) function does not estimate the smoothing parameter;
- in computational terms the pb(·) function is faster than the ps(·) function.

More details in Stasinopoulos et al. (2017).

The ps(·) and pb(·) functions can be determined from h($\mathbf{x}$)=$\mathbf{N}\boldsymbol{\gamma}$, where $\mathbf{N}$ is the incidence matrix which depends on the covariable $\mathbf{x}$ and $\boldsymbol{\gamma}$ is a parameter vector to be estimated under the matrix of B-spline bases. Further, these smoothing functions also have a quadratic penalty of the form $\lambda\boldsymbol{\gamma}^T\mathbf{G}\boldsymbol{\gamma}$, where $\mathbf{G} = \mathbf{D}^T\mathbf{D}$ is a known penalty matrix, λ is the hyperparameter that regulates the number of smoother steps necessary for adjustment and the matrix $\mathbf{D}$ is defined in (3.10).

Given this, the penalized log-likelihood function for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ can be introduced as

$$l_p(\boldsymbol{\theta}, \boldsymbol{\gamma}) = l(\boldsymbol{\theta}) - \frac{1}{2}\sum_{\xi=1}^{J}{\boldsymbol{\gamma}_\xi}^T\mathbf{G}_\xi(\boldsymbol{\lambda}_\xi)\boldsymbol{\gamma}_\xi, \tag{3.9}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma, \nu, \tau)^T$ is the vector of parameters, J is the number of smoothers or covariables, which are controlled in nonparametric form, and $\boldsymbol{\gamma}$ is a vector of penalization coefficients to be estimated. The penalty matrix $\mathbf{G}$ is defined as $\mathbf{G} = (\mathbf{D}^k)^T\mathbf{D}^k$, where the matrix $\mathbf{D}^k$ has order $(q-k) \times q$, recalling that q is the number of distinct values of the explanatory variables, which is

controlled nonparametrically. The order to be applied depends on the smoothing of the variability curve of the data. The penalization standard normally used of order $k = 2$ can be referred as

$$\boldsymbol{\gamma}^T(\mathbf{D}^2)^T\mathbf{D}^2\boldsymbol{\gamma} = (\gamma_1 - 2\gamma_2 + \gamma_3)^2 + \ldots + (\gamma_q - 2\gamma_q + \gamma_q)^2. \tag{3.10}$$

Thus, the matrix $\mathbf{D}$ has the form

$$\mathbf{D}^2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots \\ 0 & 1 & -2 & 1 & \cdots \\ 0 & 0 & 1 & -2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The asymptotic distribution of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is multivariate normal $N_{p+3}(0, K(\boldsymbol{\theta})^{-1})$ under general regularity conditions, where $K(\boldsymbol{\theta})$ is the expected information matrix. The asymptotic covariance matrix $K(\boldsymbol{\theta})^{-1}$ of $\widehat{\boldsymbol{\theta}}$ can be approximated by the inverse of the $(p+3) \times (p+3)$ observed information matrix $J(\boldsymbol{\theta})$. By doing this, the inference on the parameter vector $\boldsymbol{\theta}$ can be based on the multivariate normal distribution $N_{p+3}(0, J(\boldsymbol{\theta})^{-1})$ for $\widehat{\boldsymbol{\theta}}$ and then a $100(1-\alpha^*)\%$ asymptotic confidence interval for any parameter θ_q follows as

$$ACI_q = \left(\hat{\theta}_q - z_{\alpha^*/2}\sqrt{\widehat{J}^{q,q}}, \hat{\theta}_q + z_{\alpha^*/2}\sqrt{\widehat{J}^{q,q}}\right),$$

where $\widehat{J}^{q,q}$ denotes the qth diagonal element of the inverse of the estimated observed information matrix $J(\widehat{\boldsymbol{\theta}})^{-1}$ and $z_{\alpha^*/2}$ is the quantile $1 - \alpha^*/2$ of the standard normal distribution.

We can use LR statistics for confront with some models embedded with the OLLGIG semiparametric regression model.

3.2.1 Diagnostic tools and residual analysis

In order to assess possible influential points, an analysis of global influence may be carried from case-deletion. The case-deletion regressions with systematic components (3.4), (3.5) and (3.6) can be expressed as $\mu_l = \exp\left[\sum_\xi^J h_\xi(x_{l\xi})\right]$, $\mu_l = \exp\left[\mathbf{w}_l^T\boldsymbol{\beta} + h(x_l)\right]$and $\mu_l = \exp\left[\mathbf{w}_l^T\boldsymbol{\beta} + \sum_\xi^J h_\xi(x_{l\xi})\right]$ respectively, for $\xi = 1, \ldots, J, \quad l = 1, \ldots, n, \quad l \neq i$. The standardized norm of $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$, called the generalized Cook distance, is the first measure of the global influence defined by $GD_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^T\left[\ddot{\mathbf{L}}(\boldsymbol{\theta})\right](\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})$, where a quantity with subscript "(i)" means the original quantity with the ith observation deleted Another popular measure of the difference between $\hat{\boldsymbol{\theta}}_{(i)}$ and $\hat{\boldsymbol{\theta}}$ is the likelihood distance defined by $LD_i(\boldsymbol{\theta}) = 2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)})]$.

Once the model is chosen and fitted, the analysis of the residuals is an efficient way to check the model adequacy. For a residual analysis, we suggest working with the quantile residual (Dunn and Smyth, 1996). The qrs for the OLLGIG semiparametric regression with systematic component take the forms

$$qr_i = \Phi^{-1}\left\{\frac{\eta(y_i)^\tau}{\eta(y_i)^\tau + [1 - \eta(y_i)]^\tau}\right\}, \tag{3.11}$$

where $\eta(\cdot)$ is given in Equation (3.2) and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cdf. Atkinson (1982) suggested the construction of an envelope to have a better interpretation of the probability normal plot of the residuals.

3.3 Simulation study using different penalized smoothers

To verify the accuracy of the OLLGIG semiparametric regression MLEs with different penalties, a simulation study was performed, and also to explore the accuracy of the performance of the empirical distribution of the qrs. The response and the covariables are generated as follow: $y_i \sim \text{OLLGIG}(\mu_i, \sigma, \nu, \tau)$, $w_{i1} \sim \text{Normal}(0, 1)$, $w_{i2} \sim \text{Binomial}(1, 0.5)$ and $x_{i3} \sim \text{Uniform}(0, 7)$.

In this case, only the coefficients associated with the explanatory variables w_1 and w_2 will be analyzed, since the coefficient associated with the penalized smoothers does not have a direct explanation. Then, a graphical analysis is performed, where each of the plots presents true smooth curve defined by $h(x_{i3}) = [1 + \sin(x_{i3})]$. We consider different sample sizes ($n = 50, 100$, and 250) under three scenarios cs$(\cdot)$, ps$(\cdot)$ and pb$(\cdot)$ considering that the systematic component of the regression is $\mu_i = 0.01w_{i1} - w_{i2} + [1 + \sin(x_{i3})]$. When n increases, the adjusted curves approach the actual curve (as expected).

For these scenarios, the numeric values of the parameters are taken as: $\beta_1 = 0.01$, $\beta_2 = -1$, $\sigma = 1.5$, $\nu = 6$ and $\tau = 0.8$. Thus, for each combination of n, β_1 and β_2, $1,000$ Monte Carlo simulations are generated and for each of the samples the MLEs of the model parameters are estimated. For each replication, we evaluate the MLEs of the parameters and then, after all replications, we compute the average estimates (AEs), biases and means squared errors (MSEs). Table 3.1 provides the different systematic components for the parameter μ.

Table 3.1. Systematic components for the parameters.

Regression	Penalized smoothers	Systematic components
OLLGIG	cs$(\cdot)$	$\mu_i = \exp[\beta_1 w_{i1} + \beta_2 w_{i2} + cs(x_{i3})]$
	ps$(\cdot)$	$\mu_i = \exp[\beta_1 w_{i1} + \beta_2 w_{i2} + ps(x_{i3})]$
	pb$(\cdot)$	$\mu_i = \exp[\beta_1 w_{i1} + \beta_2 w_{i2} + pb(x_{i3})]$

From Table 3.2 you can see that the parameter EAs approach the parameters true value when n increases. Further, the biases and MSEs are small for the estimates of β_1 and β_2 even when n is small which supports that the asymptotic normal distribution provides an adequate approximation to the finite sample distribution of the MLEs.

Table 3.2. AEs, biases and MSEs for the fitted OLLGIG regression with penalized smoothers under scenarios 1[cs$(\cdot)$], 2[ps$(\cdot)$] and 3[pb$(\cdot)$].

scenario 1									
	$n = 50$			$n = 100$			$n = 250$		
Parameters	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
β_1	0.0104	0.0004	0.0052	0.0085	-0.0015	0.0023	0.0100	0.0000	0.0009
β_2	-0.9924	0.0076	0.0200	-0.9969	0.0031	0.0087	-1.0033	-0.0033	0.0035
scenario 2									
	$n = 50$			$n = 100$			$n = 250$		
Parameters	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
β_1	0.0102	0.0002	0.0051	0.0085	-0.0015	0.0022	0.0090	-0.0010	0.0009
β_2	-0.9900	0.0100	0.0199	-0.9972	0.0028	0.0089	-1.0031	-0.0031	0.0034
scenario 3									
	$n = 50$			$n = 100$			$n = 250$		
Parameters	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
β_1	0.0081	-0.0019	0.0053	0.0066	-0.0034	0.0022	0.0095	-0.0005	0.0009
β_2	-1.0068	-0.0068	0.0194	-0.9974	0.0026	0.0093	-0.9988	0.0012	0.0037

In Figure 3.1, we plot the adjusted and generated terms for the smooth functions representing the first, second and third scenarios with penalized smoothers cs$(\cdot)$, ps$(\cdot)$ and pb$(\cdot)$, respectively. For all scenarios, the generated smooth functions approximate the true curve when the sample size increases. Thus, we can conclude that the variability among the non-parametric function estimates is reduced when

n increases. We can also note that the three smoothing functions have similar performances, i.e., we can not say that anyone is better than the others. Finally, we suggest readers always to work with the three soothing functions. This same procedure is adopted in the various examples in Section 3.4 using some goodness-of-fit statistics to choose one of the three smoothing functions.

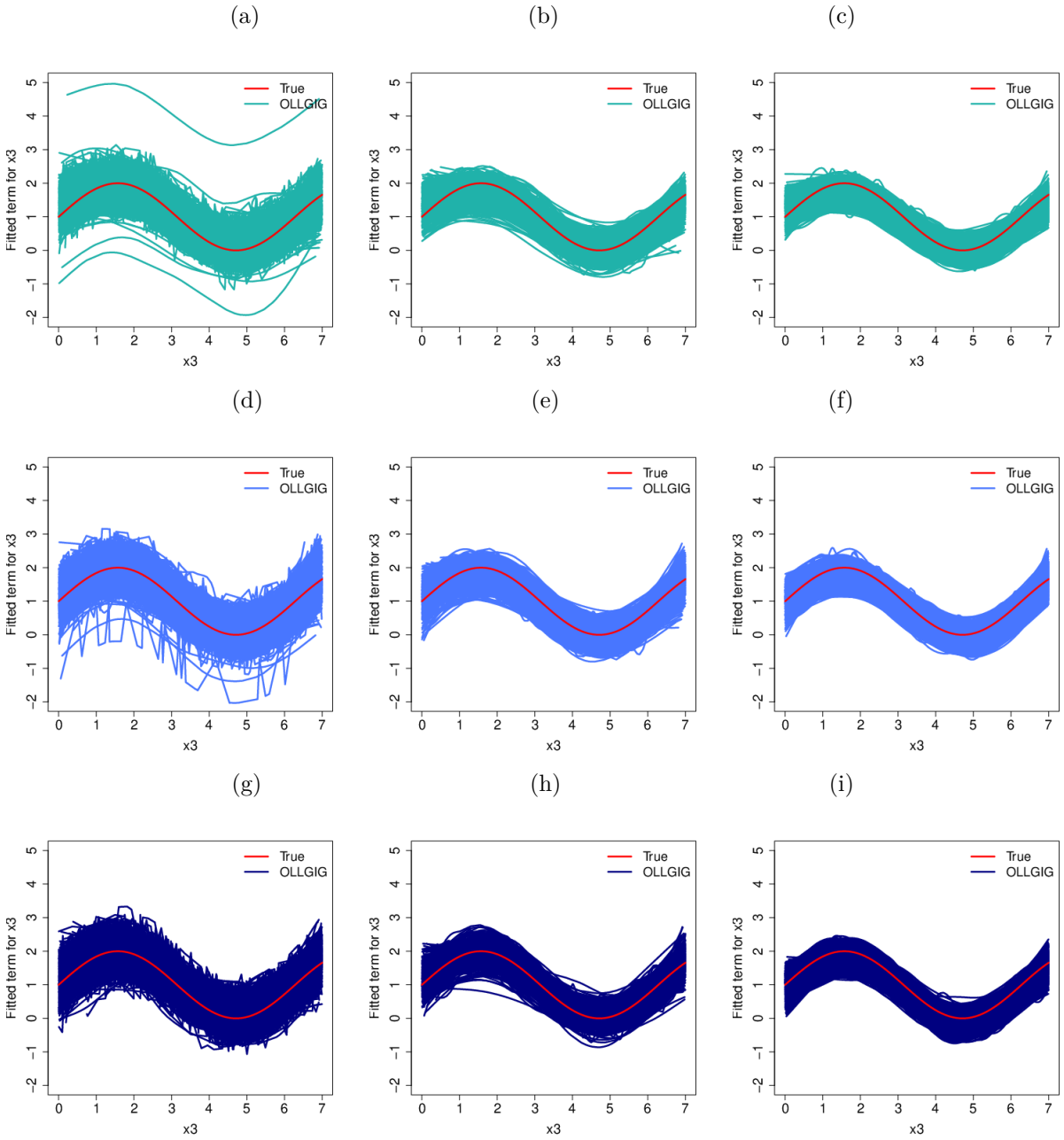

Figure 3.1. The fitted and generated terms for the smooth functions based on 1,000 simulations. The first three plots (a) $n = 50$, (b) $n = 100$ and (c) $n = 250$ are related to the scenario (1) with penalized smoother cs$(\cdot)$. The middle three plots (d) $n = 50$, (e) $n = 100$ and (f) $n = 250$ refer to the scenario (2) with penalized smoother ps$(\cdot)$. The last three plots (g) $n = 50$, (h) $n = 100$ and (i) $n = 250$ are related to the scenario (3) with penalized smoother pb$(\cdot)$.

Empirical distribution of the residuals

We have implemented a simulation study to study the empirical distribution of $(qr_i's)$ for the OLLGIG semiparametric regression model. The simulation algorithm follows the same patterns as described at the beginning of this section. We also construct the normal probability plots to assess the

degree of deviation from the normality hypothesis for the residuals. Based on the plots in Figure 3.2 representing the first, second and third scenarios, respectively, we note that the empirical distribution of these residuals agrees with the standard normal distribution for all scenarios.

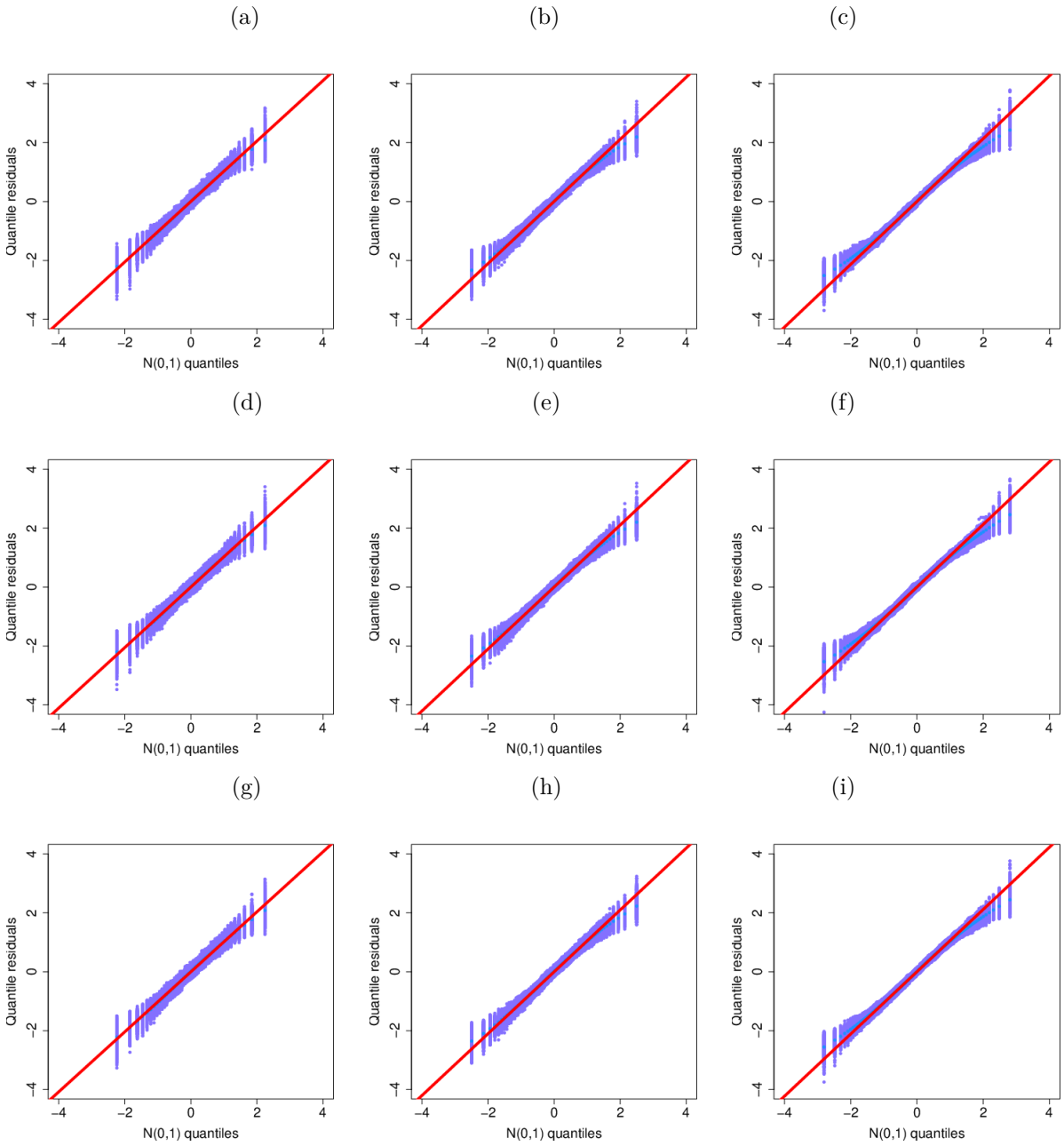

Figure 3.2. Normal probability plots for the qrs. The first three plots (a) $n = 50$, (b) $n = 100$ and (c) $n = 250$ are related to the scenario (1) with penalized smoother cs(·). The middle three plots (d) $n = 50$, (e) $n = 100$ and (f) $n = 250$ refer to the scenario (2) with penalized smoother ps(·). The last three plots (g) $n = 50$, (h) $n = 100$ and (i) $n = 250$ are related to the scenario (3) with penalized smoother pb(·).

3.4 Applications

In this section, we present three real data applications to prove empirically the flexibility of the OLLGIG additive, additive partial and semiparametric regressions with different penalized smoothers. All the computational works were implemented in the **R** software.

3.4.1 OLLGIG additive regression to climatology data

The first application is about the climatology data from the Department of Biosystems Engineering of the Luiz de Queiroz School of Agriculture, University of São Paulo (LEB-ESALQ-USP). The current data set is available at the link `http://www.leb.esalq.usp.br/leb/anos.html`. This data set was collected from March 8 to August 8, 2019. We consider the OLLGIG additive regression to explore the influence of the covariables (global radiation, relative humidity and maximum wind) in the evaporation (response variable). Then, the variables considered for this application are:

- y_i: Evaporation (mm);
- x_{i1}: Global radiation (cal/cm^2);
- x_{i2}: Relative humidity (%);
- x_{i3}: Maximum wind (m/s), for $i = 1, \ldots, 154$.

In Table 3.3 the MLEs are shown, their standard errors (SEs) in parentheses, the values of the Akaike information criterion (AIC) and global deviation (GD). The fitted model is better suited when the values of these criteria are small. The lower values of the two statistics in this table support that the OLLGIG distribution would be right for modeling these data.

Table 3.3. MLEs and SEs of the model parameters for climatology data.

Model	$\log(\mu)$	$\log(\sigma)$	ν	τ	AIC	GD
OLLGIG	**1.2780**	**-1.9504**	**27.8990**	**0.2875**	**482.7119**	**474.7119**
	(0.0226)	**(0.0554)**	**(9.8170)**	**(0.0193)**		
GIG	1.3176	-1.0429	3.3460	1	491.1583	487.1583
	(0.0269)	(0.1295)	(5.3590)	(-)		
IG	1.3178	-1.7319	-0.5	1	492.6715	486.6715
	(0.0276)	(0.0569)	(-)	(-)		

The proposed distribution is associate with two sub-models using LR statistics in Table 3.4. The figures in this table, specially the p-values, reveal that the OLLGIG model gives a better fit to these data than its two sub-models. Plots of the fitted OLLGIG, GIG and IG densities are displayed in Figure 3.3(b) to assess the appropriateness of the models. Plots of the estimated cumulative and the empirical distributions are exposed in Figure 3.3(c). They reveal that the OLLGIG distribution offers a efficient fit to the current data, thus capturing a slight bimodality with left asymmetry.

Table 3.4. LR tests for climatology data.

Models	Hypotheses	Statistic w	p-value
OLLGIG vs GIG	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	11.9596	0.0005
OLLGIG vs IG	$H_0 : \tau = 1$ and $\nu = -0.5$ vs $H_1 : H_0$ is false	12.4464	0.0019

Regression analysis with systematic components

We note in Figure 3.4 that there is a nonlinear relationship between the response variable and each of the covariables x_1, x_2 and x_3. Thus, the OLLGIG additive regression model is a good option for modeling these data. The systematic components for the parameter μ in Table 3.5 represent the OLLGIG, GIG and IG additive regressions with penalized smoothers for the explanatory variables x_1, x_2 and x_3. The generalized Akaike information criterion (GAIC) measure is adopted for model selection (Rigby and Stasinopoulos, 2005) because smoothing terms are included in the systematic components. The measures of this statistic are displayed in Table 3.5 to verify the adequacy of all fitted models. They show that the fitted OLLGIG additive regression with pb$(\cdot)$ smoother has the lowest measure for the

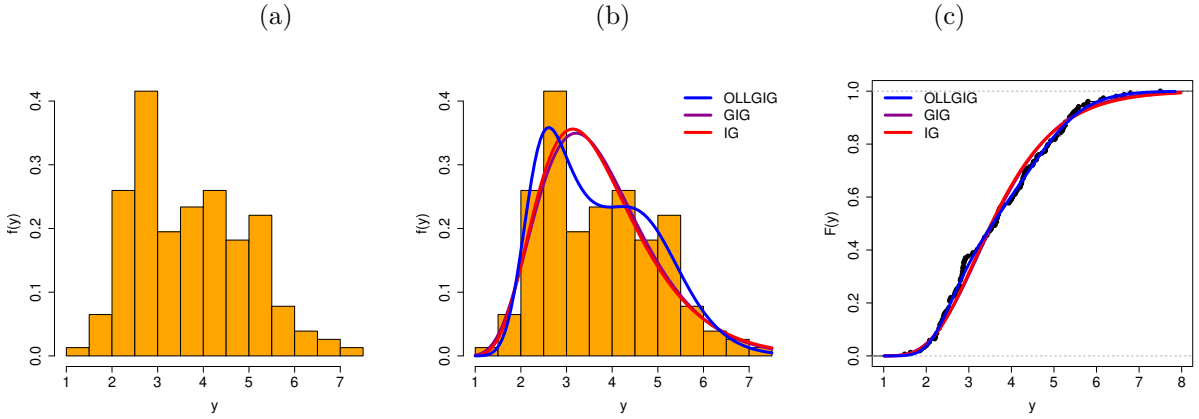

Figure 3.3. (a) Histogram of the evaporation variable. (b) Estimated OLLGIG, GIG and IG densities for climatology data. (c) Estimated cdf of the OLLGIG, GIG and IG distributions and the empirical cdf.

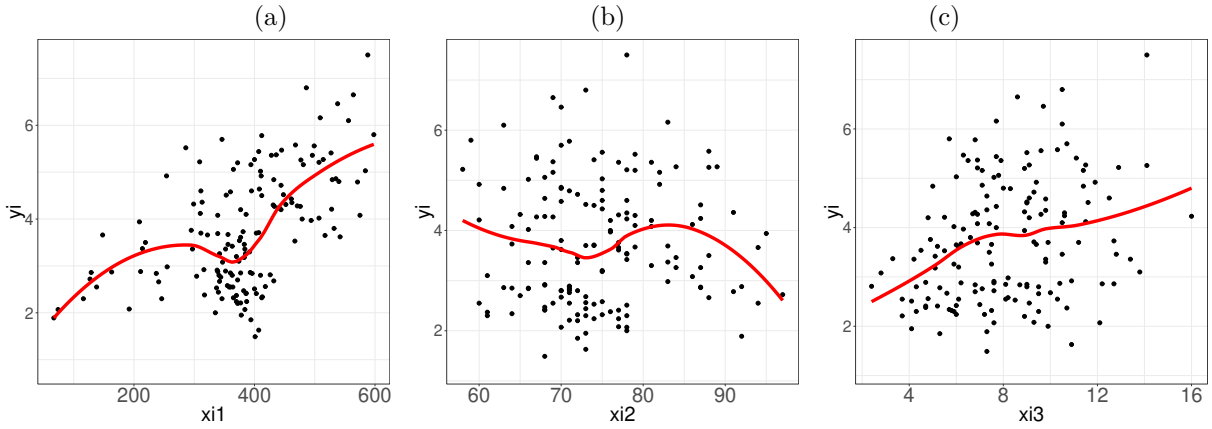

Figure 3.4. Dispersion diagrams for climatology data. (a) y versus x_1. (a) y versus x_2. (a) y versus x_3.

Table 3.5. Systematic components of the OLLGIG, GIG and IG additive regressions and goodness-of-fit measures for climatology data.

Model	systematic structures	GAIC
OLLGIG	$\mu_i = \exp[\beta_0 + cs(x_{i1}) + cs(x_{i2}) + cs(x_{i3})]$	402.3305
GIG	$\mu_i = \exp[\beta_0 + cs(x_{i1}) + cs(x_{i2}) + cs(x_{i3})]$	407.7017
IG	$\mu_i = \exp[\beta_0 + cs(x_{i1}) + cs(x_{i2}) + cs(x_{i3})]$	413.6021
OLLGIG	$\mu_i = \exp[\beta_0 + ps(x_{i1}) + ps(x_{i2}) + ps(x_{i3})]$	408.9545
GIG	$\mu_i = \exp[\beta_0 + ps(x_{i1}) + ps(x_{i2}) + ps(x_{i3})]$	413.9161
IG	$\mu_i = \exp[\beta_0 + ps(x_{i1}) + ps(x_{i2}) + ps(x_{i3})]$	419.9604
OLLGIG	$\mu_i = \exp[\beta_0 + pb(x_{i1}) + pb(x_{i2}) + pb(x_{i3})]$	**401.8478**
GIG	$\mu_i = \exp[\beta_0 + pb(x_{i1}) + pb(x_{i2}) + pb(x_{i3})]$	405.1313
IG	$\mu_i = \exp[\beta_0 + pb(x_{i1}) + pb(x_{i2}) + pb(x_{i3})]$	410.8665

GAIC statistic among the fitted regressions. The MLEs of the model parameters listed in Table 3.6 are evaluated. Additional interpretations for this regression will be made at the end of this subsection. Table 3.7 compares the proposed distribution with two sub-models via LR statistics, where the p-values support that the OLLGIG additive regression with pb$(\cdot)$ provides a conducive fit to the current data than the null models. It was calculated the case-deletion measures $GD_i(\boldsymbol{\theta})$ and $LD_i(\boldsymbol{\theta})$ defined in Subsection 3.2.1. The results of such influence measure index plots are presented in Figure 3.5. The plots reveal that the cases ♯61, ♯89, ♯122 and ♯146 are possible influential observations. We perform the residual analysis by plotting in Figure 3.5(c) the qrs (see Subsection 3.2.1) against the index of the observations. Figure

Table 3.6. MLEs, SEs and p-values for the OLLGIG additive regression with pb(·) fitted to climatology data.

Parameter	Estimate	SE	p-value
β_0	-0.2576	0.1694	0.1305
$\log(\sigma)$	0.0169	0.2495	
ν	0.4622	0.4266	
τ	4.3021	0.6548	

Table 3.7. LR tests for comparing regressions.

Regressions	Hypotheses	Statistic w	p-value
OLLGIG pb(·) vs GIG pb(·)	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	6.0701	0.0244
OLLGIG pb(·) vs IG pb(·)	$H_0 : \tau = 1$ and $\nu = -0.5$ vs $H_1 : H_0$ is false	14.6503	0.0018

(a) (b) (c) (d)

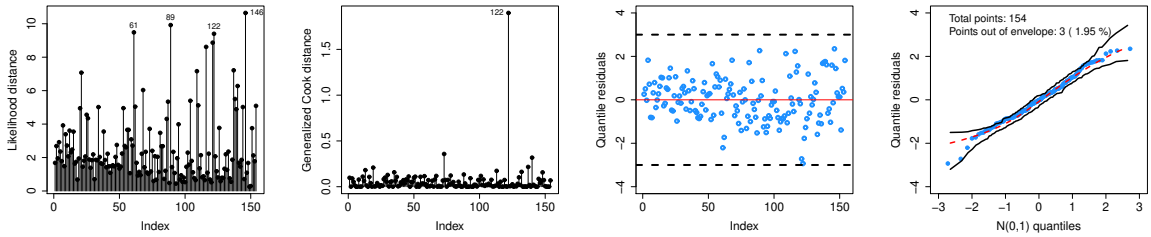

Figure 3.5. Index plots for $\boldsymbol{\theta}$: (a) $LD_i(\boldsymbol{\theta})$ (likelihood distance) and (b) $GD_i(\boldsymbol{\theta})$ (generalized Cook's distance). (c) Residual analysis of the OLLGIG additive regression with pb(·) smoother fitted to the climatology data. (d) Normal probability plot for the qrs with envelope.

3.5(d) gives the normal probability plot with generated envelope. So, the OLLGIG additive regression with pb(·) it is very appropriate for this data, although it has three observations out of the envelope, yet the percentage is less than 5%.

Final interpretations

Figure 3.6 shows the estimation of the nonlinear effects. The horizontal axis in Figure 3.6(a) refers to the values of the covariable x_1 and the vertical axis gives the contribution of the penalized smoother pb(·) for the adjusted values of the response variable (evaporation in *mm*). Note that the global radiation has a nonlinear relation with evaporation, such that:

- for days with global radiation (x_1) of up to 300 cal/cm^2 (approximately), there is an increase in evaporation;
- for days with global radiation between 300 cal/cm^2 and 380 cal/cm^2, there is a reduction of the evaporation;
- for global radiation values near 380 cal/cm^2, the evaporation is increasing.

The effect of humidity (in %) also has a nonlinear effect to the evaporation [see Figure 3.6(b)]. Further, for days having relative humidity (x_2) up to 70% (approximately), there is constant evaporation, but for days with relative humidity greater than 70% (approximately), the evaporation increases. Further, the maximum wind speed (covariable x_3) has a nonlinear effect on evaporation. For days with maximum wind speed between 2 m/s and 10 m/s (approximately), there is a rising evaporation rate, while on days with wind speeds greater than 10 m/s (approximately), the increase of evaporation is less pronounced, as can be noted in Figure 3.6(c).

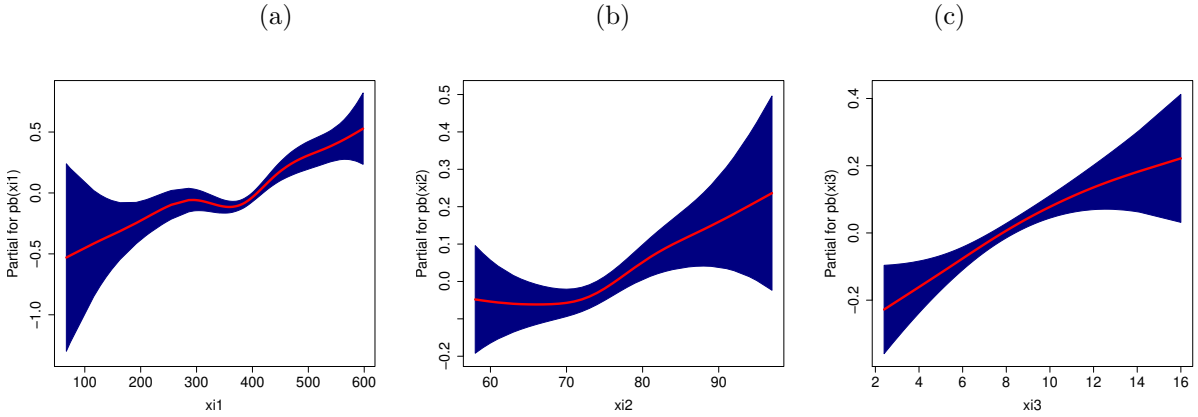

Figure 3.6. Shapes of the penalized smoothers pb(·) for the covariables (a) x_1, (b) x_2 and (c) x_3 using the OLLGIG additive regression.

3.4.2 OLLGIG additive partial regression fitted to ethanol data

This application is about the fuel ethanol burned in one cylinder engine. For various configurations of compression ratio and engine equivalency, nitrogen oxides (NOx) emissions were recorded. The ethanol data frame contains 88 sets of measurements for variables from an experiment in which ethanol was burned in a single cylinder automobile test engine. For more details about the data, see Brinkman (1981). We consider the OLLGIG additive partial regression given in Equation (3.5) in comparison to the GIG and IG additive partial regressions with three types of penalized smoothers in the linear predictor, namely: cs(·), ps(·) and pb(·).

The variables in this study are:

- y_i: NOx (concentration of nitrogen oxides (NO and NO2) in micrograms/J);
- w_{i1}: the compression ratio of the engine;
- x_{i2}: equivalence ratio, a measure of the richness of the air and ethanol fuel mixture (for $i = 1, \ldots, 88$).

Table 3.8 lists the MLEs of the parameters, their SEs and the AIC and GD measures for the OLLGIG, GIG and IG distributions. The statistics in this table reveal that the OLLGIG distribution presents the lowest values among those of all fitted distributions. So, it could be designated as the best distribution for current data.

Table 3.8. MLEs and SEs (in parentheses) of the model parameters for ethanol data.

Model	$\log(\mu)$	$\log(\sigma)$	ν	τ	AIC	GD
OLLGIG	**0.5341**	**-1.1017**	**22.8500**	**0.1951**	**250.8621**	**242.8621**
	(0.0662)	**(0.9342)**	**(9.8800)**	**(0.0656)**		
GIG	0.6717	-0.1787	1.7900	1	262.7169	256.7169
	(0.0678)	(0.2475)	(1.0310)	(-)		
IG	0.6716	-0.6436	-0.5	1	265.1399	261.1399
	(0.0783)	(0.0754)	(-)	(-)		

The LR statistics to confront nested distributions are reported in Table 3.9. Clearly, the OLLGIG distribution outperforms the GIG and IG distributions. The plots of the fitted OLLGIG, GIG and IG densities are exposed in Figure 3.7(a). It is clear that the histogram of the data has a bimodal shape and that the estimated OLLGIG density provides the plus approximate fit to the histogram. The plots for the GIG and IG densities can not have this shape. Further, the plots of the fitted OLLGIG,

Table 3.9. LR tests for ethanol data

Models	Hypothesis	Statistic w	p-value
OLLGIG vs GIG	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	13.8548	<0.001
OLLGIG vs IG	$H_0 : \tau = 1$ and $\nu = -0.5$ vs $H_1 : H_0$ is false	18.2778	<0.001

GIG and IG cdf and the empirical cdf are exposed in Figure 3.7(b). They also pointing that the wider distribution features a appropriate fit to these data. Thus, the OLLGIG distribution is a good choice for modeling the current data.

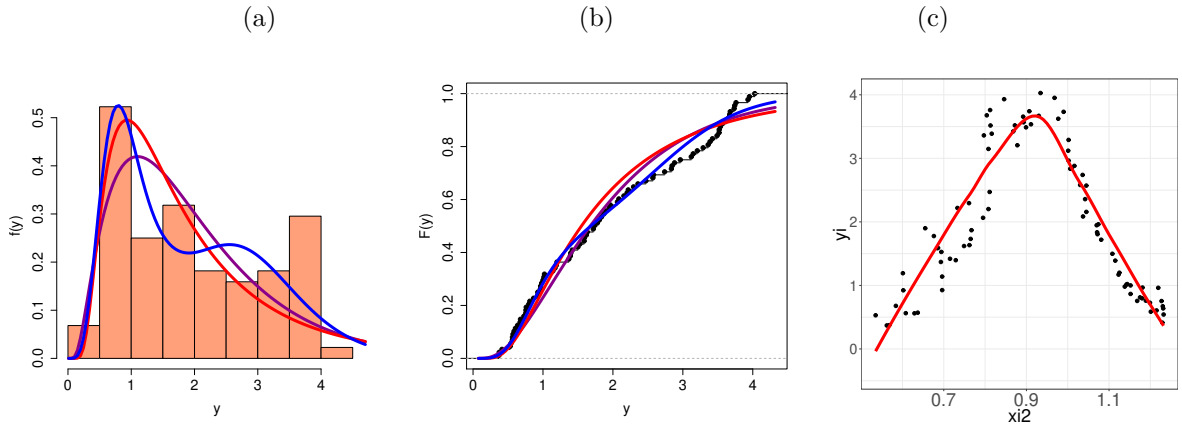

Figure 3.7. (a) Estimated OLLGIG, GIG and IG densities for ethanol data. (b) Estimated cumulative functions of the OLLGIG, GIG and IG distributions and the empirical cdf for ethanol data. (c) Scatter diagram: emission of NOx versus air/ethanol mix.

Regression analysis with systematic components

We can note from Figure 3.7(c) that there is a nonlinear effect between the response variable y and the explanatory variable x_2. So, we adopt the OLLGIG additive partial regression with different penalized smoothers. For the OLLGIG, GIG and IG additive partial regressions, the systematic components for the parameter μ by taking the nonlinear effect in the explanatory variable x_2 are given in Table 3.10. The values of the GAIC statistic for the nine fitted regressions are reported in Table 3.10. Based

Table 3.10. Additive partial regressions and GAIC for some regressions fitted to the ethanol data.

Model	systematic structures	GAIC
OLLGIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + cs(x_{i2})]$	36.2462
GIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + cs(x_{i2})]$	39.4854
IG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + cs(x_{i2})]$	91.4063
OLLGIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + ps(x_{i2})]$	39.3340
GIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + ps(x_{i2})]$	41.0862
IG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + ps(x_{i2})]$	92.2585
OLLGIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + pb(x_{i2})]$	**35.0047**
GIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + pb(x_{i2})]$	37.6757
IG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + pb(x_{i2})]$	91.0999

on these numerical results, the GAIC measure for the OLLGIG additive partial regression with penalized smoother pb$(\cdot)$ is the smallest among those of the nine fitted regressions. Hence, the proposed regression can be chosen as the best model for the current data.

Table 3.11 gives the MLEs, SEs and p-values of the model parameters. We can note that the linear (w_1) and non-linear (x_2) effects are statistically significant at 5%. Thus, an interpretation of the linear effect is that, as the compression ratio of the motor increases, so does the NOx emission. The interpretation of the non-linear effect is addressed at the end of this application. For comparing the

Table 3.11. MLEs, SEs and p-values for the OLLGIG additive partial regression with pb(·) fitted to ethanol data.

Parameter	Estimate	SE	p-value
β_0	-1.3744	0.0791	<0.001
β_1	0.0261	0.0041	<0.001
$\log(\sigma)$	20.3111	0.0051	
ν	4.4625	0.5352	
τ	3.3840	0.2694	

regressions, we consider LR statistics and formal tests. The values of the LR statistics for testing two sub-models of the OLLGIG additive partial regression are given in Table 3.12. These values yield favorable indications for the OLLGIG additive partial regression with pb(·) penalized smoother. The case-deletion

Table 3.12. LR statistics for testing some regressions.

Models	Hypotheses	Statistic w	p-value
OLLGIG pb vs GIG pb	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	4.2999	0.0281
OLLGIG pb vs IG pb	$H_0 : \tau = 1$ and $\nu = -0.5$ vs $H_1 : H_0$ is false	60.7959	<0.001

measures $GD_i(\boldsymbol{\theta})$ and $LD_i(\boldsymbol{\theta})$ are presented in the plots of Figure 3.8(a) and 3.8(b), which show that the cases ♯14, ♯24, ♯38 and ♯88 are likely influential observations. The plot of the qrs versus adjusted values is given in Figure 3.8(c) for detecting possible outliers in the OLLGIG additive partial regression with pb(·) smoother. We note that the residuals have a random behavior and there is no observation outside the range $[-3, 3]$. Figure 3.8(d) displays the normal probability plot for the qrs with the simulated envelope, which shows the good adequacy of the fitted regression. Finally, is presented the estimation of

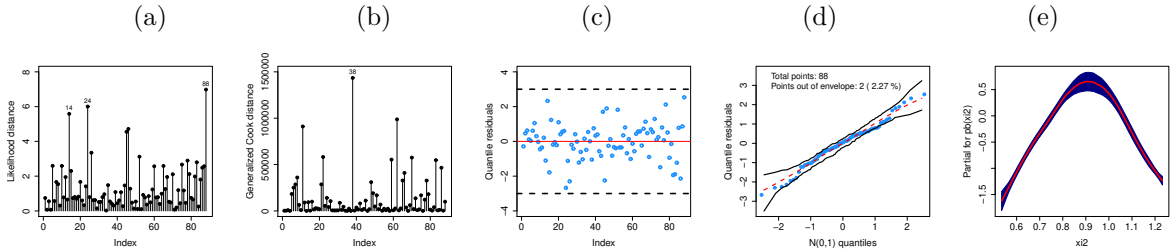

Figure 3.8. (a) $LD_i(\boldsymbol{\theta})$ (likelihood distance). (b) $GD_i(\boldsymbol{\theta})$ (generalized Cook's distance).(c) Residual analysis of the OLLGIG additive partial regression with pb(·) smoother fitted to the ethanol data. (d) Normal probability plot for the qrs with envelope. (e) Shape of the penalized smoothers pb(·) for the covariable x_2.

the nonlinear effect in Figure 3.8(e). In the horizontal axis, we have the values of the covariant x_2 and in the vertical axis the contribution of the penalized smoother pb(·) to the adjusted values of the NOx emission. The effect of the air/ethanol mix is nonlinear in relation to the NOx emission (as expected). Further, for values of x_2 around 0.9, there is an increase in NOx emission which also presents a greater variability, but from 0.9, the equivalence ratio x_2 decreases with little variability of NOx emission.

3.4.3 OLLGIG semiparametric regression fitted to air quality data

The application refers to the air quality data (*airquality*) available in the **R** software. For this analysis, the lines with missing information were omitted. The data are the daily air quality readings (from May 1 to September 30, 1973) obtained from the New York State, Department of Environmental Conservation (ozone data) and the U.S. National Weather Service (meteorological data) (more details see Tukey (1983)). In this application, is it used the OLLGIG semiparametric regression and compare it with

the GIG and IG sub-models, where the systematic component is given in Equation (3.6) to describe the relation between the air quality and the other covariables. We also consider (as in the first application) the penalized smoothers cs(·), ps(·) and pb(·) in the linear predictors. The data are:

- y_i: average ozone concentration in parts per billion from 1:00 to 3:00 p.m. on Roosevelt Island;
- w_{i1}: the explanatory variable month, considered as a factor with five levels (May, June, July, August and September);
- x_{i2}: solar radiation in Langleys in the frequency range from 4000-7700 Angstroms from 8:00 a.m. to 12:00 noon in Central Park;
- x_{i3}: maximum daily temperature in degrees Fahrenheit at La Guardia Airport, $i = 1, \ldots, 111$. .

Figure 3.9 shows that there is a nonlinear relationship between the response variable and each of the covariables x_2 and x_3. Then, we adopt the OLLGIG semiparametric regression with different penalized smoothers to analyze these data. Table 3.13 presents the OLLGIG, GIG and IG semiparametric regressions with different systematic components with nonlinear effects in the explanatory variables x_2 and x_3. The values of the GAIC measure for the nine fitted regressions are listed in Table 3.13. The OLLGIG

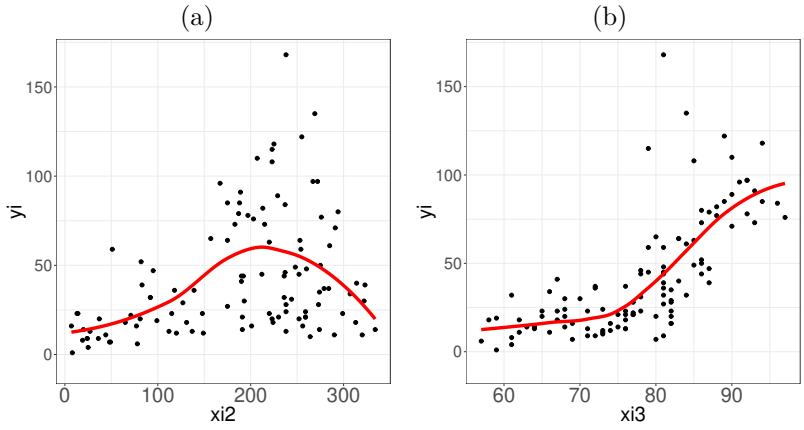

Figure 3.9. Scatter diagram: (a) y_i versus x_{i2}. (b) y_i versus x_{i3}.

Table 3.13. Semiparametric Regressions and GAIC statistic from the fitted regressions to the air quality data.

Model	systematic components	GAIC
OLLGIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + cs(x_{i2}) + cs(x_{i3})]$	936.3173
GIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + cs(x_{i2}) + cs(x_{i3})]$	940.2850
IG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + cs(x_{i2}) + cs(x_{i3})]$	1019.6003
OLLGIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + ps(x_{i2}) + ps(x_{i3})]$	**934.7905**
GIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + ps(x_{i2}) + ps(x_{i3})]$	939.3618
IG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + ps(x_{i2}) + ps(x_{i3})]$	1017.3683
OLLGIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + pb(x_{i2}) + pb(x_{i3})]$	941.1109
GIG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + pb(x_{i2}) + pb(x_{i3})]$	941.7588
IG	$\mu_i = \exp[\beta_0 + \beta_1 w_{i1} + pb(x_{i2}) + pb(x_{i3})]$	1018.3338

semiparametric regression with ps(·) smoother has the smallest GAIC among those of the nine fitted regressions, and then it can be indicated as the best model. Table 3.14 gives the MLEs, SEs and p-values of the model parameters. For the 5% significant level, the explanatory variable w_1 is significant. Since the values of the estimates are negative, there is a strong evidence in June and September and a beginning of a lower average level of ozone in May. The values of the LR statistics for testing two sub-models of

Table 3.14. MLEs, SEs and p-values for the fitted semiparametric OLLGIG regression with ps$(\cdot)$ to the air quality data.

Parameter	Estimate	SE	p-value
β_0	-0.6978	0.5333	0.1938
β_{11}	-0.3783	0.1737	0.0318
β_{12}	-0.1439	0.1715	0.4033
β_{13}	-0.1056	0.1716	0.5395
β_{14}	-0.3329	0.1418	0.0209
$\log(\sigma)$	3.7104	0.3348	
ν	0.2315	0.0251	
τ	6.9203	0.4931	

the OLLGIG semiparametric regression with the ps$(\cdot)$ smoother are reported in Table 3.15, which yield favorable indications for the wider semiparametric regression. Generalized Cook's distance $GD_i(\boldsymbol{\theta})$ and

Table 3.15. LR tests for some semiparametric regressions.

Models	Hypotheses	Statistic w	p-value
OLLGIG ps vs GIG ps	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	6.5713	0.0104
OLLGIG ps vs IG ps	$H_0 : \tau = 1$ and $\nu = -0.5$ vs $H_1 : H_0$ is false	86.5778	<0.001

likelihood distance $LD_i(\boldsymbol{\theta})$ are displayed in Figure 3.10. These plots show that the cases ♯23 and ♯77 are possible influential observations. On the other hand, the plot of the qrs versus the fitted is explicit in

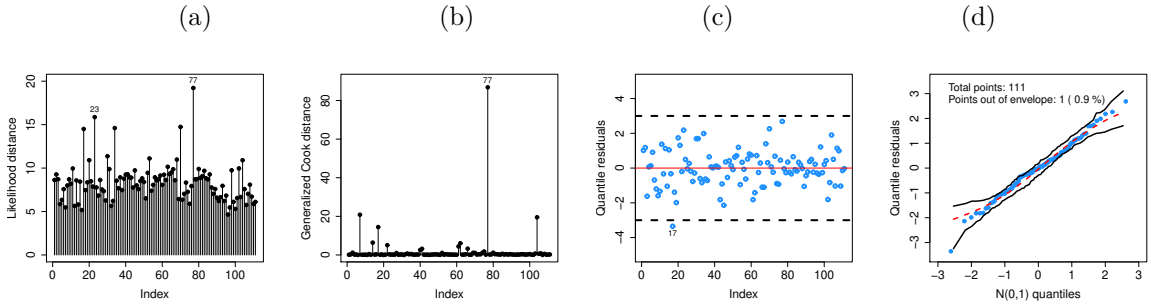

Figure 3.10. Index plots for $\boldsymbol{\theta}$: (a) $LD_i(\boldsymbol{\theta})$ (likelihood distance) and (b) $GD_i(\boldsymbol{\theta})$ (generalized Cook's distance). (c) Residual analysis of the fitted OLLGIG semiparamteric regression to the current data. (d) Normal probability plot for the qrs with envelope.

Figure 3.10(c). It is clear a random performance of the residuals around the x-axis and that the observation ♯17 is outside the range $[-3, 3]$. We verify the quality of the adjustment range of the OLLGIG semiparametric regression by the normal probability plot for the rqs with the simulated envelope given in Figure 3.10(d). This plot supports the good fit of the OLLGIG semiparametric regression with ps$(\cdot)$ to the current data. The values of the covariables x_2 and x_3 are expressed in the horizontal axis of Figure 3.11 and the contribution of the penalized smoother ps$(\cdot)$ in each of these covariables in the vertical axis. We note that the effects of solar radiation and temperature are nonlinear as expected. We have two conclusions:

- The penalized smoother for x_2 as noted in Figure 3.11(a) presents an increasing period of median ozone incidence and the decay of the adjusted curve from 240 (approximately). In relation to the variability remained constant, only above the level of solar radiation around 300 occurred an increase in the variability of the median incidence of ozone.

- The functional form of the covariable x_3 in Figure 3.11(b) shows a continuous increase in the median incidence of ozone in relation to the temperature up to around 95 degrees Fahrenheit, thus tending

to decrease the adjusted curve. Further, there is a considerable increase in the variability of median ozone incidence when the temperature is above 95 degrees Fahrenheit.

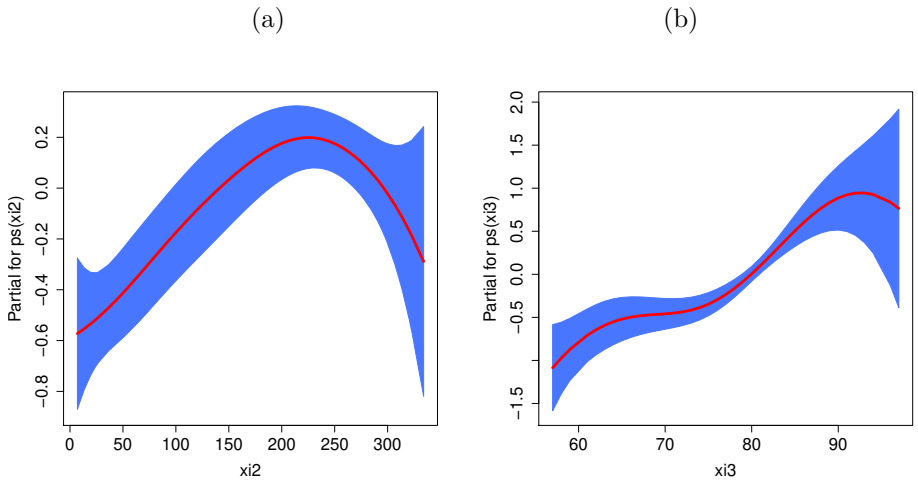

Figure 3.11. Shapes of the penalized smoothers ps(·) for the covariables x_2 and x_3 via the OLLGIG semiparametric regression model.

3.5 Concluding Remarks

This paper presents the additive, partial additive and semiparametric regression models under a distribution, called the *odd log-logistic generalized inverse Gaussian* (OLLGIG), which are very flexible for both unimodal and bimodal data. The proposed regressions include as embedded models the generalized inverse Gaussian and inverse Gaussian regressions in addition to the systematic components with three types of penalized smoothers. The proposed regressions extend some existing additive, additive partial and semiparametric regressions and they can be valuable additions for search line in regression models and extensions. The maximum penalized likelihood method is detailed to estimate the model parameters. The sensitivity of penalized maximum likelihood estimates of adjusted regressions using quantile residuals was also discussed. The versatility of the proposed regressions is proved empirically by through of three applications to climatology, ethanol and air quality data.

References

Atkinson, A. (1982). The simulation of generalized inverse gaussian and hyperbolic random variables. *SIAM Journal on Scientific and Statistical Computing*, 3(4):502–515.

Brinkman, N. D. (1981). Ethanol fuel—single—cylinder engine study of efficiency and exhaust emissions. *SAE Transactions*, pages 1410–1424.

Del Giudice, V., Manganelli, B., and De Paola, P. (2015). Spline smoothing for estimating hedonic housing price models. In *International Conference on Computational Science and Its Applications*, pages 210–219. Springer.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, pages 89–102.

Etienne, X. L., Ferrara, G., and Mugabe, D. (2019). How efficient is maize production among smallholder farmers in zimbabwe? a comparison of semiparametric and parametric frontier efficiency analyses. *Applied Economics*, 51(26):2855–2871.

Fan, S. and Hyndman, R. J. (2011). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, 27(1):134–141.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.

Green, P. J. and Yandell, B. S. (1985). Semi-parametric generalized linear models. In *Generalized Linear Models*, pages 44–55. Springer.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman and Hall.

Hudson, I. L., Kim, S. W., and Keatley, M. R. (2010). Climatic influences on the flowering phenology of four eucalypts: a gamlss approach. In *Phenological Research*, pages 209–228. Springer.

Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution*. Springer Science & Business Media, New York.

Lebotsa, M. E., Sigauke, C., Bere, A., Fildes, R., and Boylan, J. E. (2018). Short term electricity demand forecasting using partially linear additive quantile regression with an application to the unit commitment problem. *Applied Energy*, 222:104–118.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Ruppert, D., Wand, M., and Carroll, R. (2003). Semiparametric regression.,(cambridge university press: Cambridge, uk.).

Souza Vasconcelos, J. C., Cordeiro, G. M., Ortega, E. M., and Araújo, E. G. (2019). The new odd log-logistic generalized inverse gaussian regression model. *Journal of Probability and Statistics*, 2019.

Stasinopoulos, D. M., Rigby, R. A., et al. (2007). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46.

Stasinopoulos, M., Rigby, B., and Akantziliotou, C. (2008). Instructions on how to use the gamlss package in r second edition.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible regression and smoothing: using GAMLSS in R*. Chapman and Hall/CRC.

Tukey, P. A. (1983). Graphical methods. In *Proceedings of Symposia in Applied Mathematics*, volume 28, pages 8–48.

4 A RANDOM EFFECT REGRESSION BASED ON THE ODD LOG-LOGISTIC GENERALIZED INVERSE GAUSSIAN DISTRIBUTION

Abstract: A random effect regression is defined to model correlated data. The maximum likelihood is adopted to estimate the parameters and various simulations are performed for correlated data. A type of residuals for the new regression is proposed whose empirical distribution is close to normal. The usefulness of the regression is verified based on the average price per hectare of bare land in 10 cities in the state of São Paulo (Brazil).
Keywords: Bimodal data; generalized inverse Gaussian; hectare price data; regression model; simulation study.

4.1 Introduction

Many studies in the fields of public health, economics, agronomy, medicine, biology and the social sciences, among others, involve repeated observations of a response variable. The expression "repeated measures" is used to designate measures obtained for the same variable or in the same experimental unit on more than one occasion; see Diggle (1988); Crowder and Hand (1990). Various experimental designs with repeated measures are common, such as split-plot, crossover and longitudinal. These types of investigations are referred to as correlated data studies, and they play a fundamental role in the analysis of results where it is possible to characterize alterations in the characteristics of an individual by associating these variations with a set of covariables. Due to their nature the repeated measures have a correlation structure that plays an important role in the analysis of these types of data. Besides, the distribution of the response variable can present asymmetry or bimodality.

Recently, some works in this area were developed. For example, Muniz-Terrera et al. (2016) developed random effect parametric and nonparametric regressions for analyze cognitive test data, Coupé (2018) reported advances in statistical modeling in linguistics based in linear mixed-effects regressions, Ho et al. (2019) presented an analysis of microbiome relative abundance data using a zero-inflated beta GAMLSS and meta-analysis across studies using random effects models, Hashimoto et al. (2019) introduced a random effect log-Burr XII regression and Dirmeier et al. (2020) presented host factor prioritization for pan-viral genetic perturbation screens using random intercept models and network propagation.

Figure 4.1 display the average price per hectare of bare land in 10 cities in the state of São Paulo, Brazil, where bare land consists of the soil and its surface with the respective vegetation, such as forest or pasture. These data were obtained from the website of the Institute of Agricultural Economics (IEA) and the Office to Coordinate Integral Technical Assistance (CATI) referring to 2015. In this case, each city is interpreted as a group, and the data within each city are correlated, while between cities they are considered independent.

So, to analyze correlated data in the presence of bimodality and asymmetry, and based on the studies described, it is introduced a regression with normal random intercepts based on the *odd log-logistic generalized inverse Gaussian* (OLLGIG) distribution for the purpose of considering the possible presence of heterogeneity among the cities.

The remainder of this paper is divided into four sections. In Section 4.2, the random effect OLLGIG regression is defined. In Section 4.3, the maximum likelihood estimators (MLEs) are obtained via numerical integration method, some simulations are performed and the quantile residuals are defined. In Section 4.4, a real data set is analyzed for illustrative purposes. Finally, some conclusions are offered in Section 4.5.

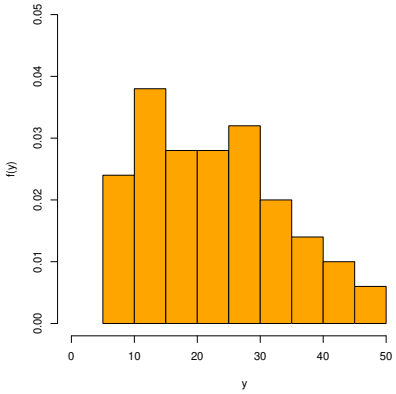

Figure 4.1. Histograma of the average price per hectare of raw land.

4.2 The random effect OLLGIG regression

Many generalized *inverse Gaussian* (IG) distributions aim to provide better fits to certain data sets than the traditional two or three parameter IG models. The *generalized inverse Gaussian* (GIG) distribution with three parameters (Jørgensen, 1982) presents several properties and applications of this distribution. Good properties and being more flexible, the GIG distribution still did not have the appropriateness to model bimodal data. In this context, Souza Vasconcelos et al. (2019) introduced a new generalization of the GIG distribution called the *odd log-logistic generalized inverse Gaussian* (OLLGIG) distribution with four parameters. The most important feature of this OLLGIG distribution is that it can model bimodal data.

The cumulative distribution function (cdf) of the OLLGIG model is

$$F(y) = F(y;\mu,\sigma,\nu,\tau) = \frac{G_{\mu,\sigma,\nu}(y)^{\tau}}{G_{\mu,\sigma,\nu}(y)^{\tau} + [1 - G_{\mu,\sigma,\nu}(y)]^{\tau}}, \qquad y > 0, \tag{4.1}$$

where

$$G_{\mu,\sigma,\nu}(y) = \int_0^y \left(\frac{b}{\mu}\right)^{\nu} \frac{t^{\nu-1}}{2K_{\nu}(\sigma^{-2})} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{b\,t}{\mu} + \frac{\mu}{b\,t}\right)\right] dt \tag{4.2}$$

is the cdf of the GIG distribution, $\mu > 0$ represents its mean, $\sigma > 0$ is a scale parameter, and $\nu \in \mathbb{R}$ and $\tau > 0$ are shape parameters,

$$b = \frac{K_{\nu+1}(\sigma^{-2})}{K_{\nu}(\sigma^{-2})} \qquad \text{and} \qquad K_{\nu}(t) = \frac{1}{2}\int_0^{\infty} y^{\nu-1} \exp\left[-\frac{1}{2}t\left(u + u^{-1}\right)\right] du \tag{4.3}$$

is the modified Bessel function of the third kind and index ν. Clearly, $G_{\mu,\sigma,\nu}(y)$ follows from (4.1) if $\tau = 1$. Further details and properties of the GIG distribution can be found in Jørgensen (1982).

If $\eta(y) = G_{\mu,\sigma,\nu}(y)$, the OLLGIG density function (for $y > 0$) can be expressed as

$$\begin{aligned} f(y) \quad = \quad & f(y;\mu,\sigma,\nu,\tau) = \left(\frac{b}{\mu}\right)^{\nu} \frac{\tau\, y^{\nu-1}}{2K_{\nu}\left(\sigma^{-2}\right)} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{by}{\mu} + \frac{\mu}{b\,y}\right)\right] \times \\ & \{\eta(y)[1-\eta(y)]\}^{\tau-1} \left\{\eta(y)^{\tau} + [1-\eta(y)]^{\tau}\right\}^{-2}. \end{aligned} \tag{4.4}$$

Figure 4.2 displays plots of the density function 4.4 for some parameter values thus showing that the OLLGIG distribution could be very flexible for modeling bimodal data.

The quantile function (qf) corresponding to (4.1) has the simple form

$$y = Q_{GIG}\left(\frac{u^{1/\tau}}{u^{1/\tau} + [1-u]^{1/\tau}}\right), \tag{4.5}$$

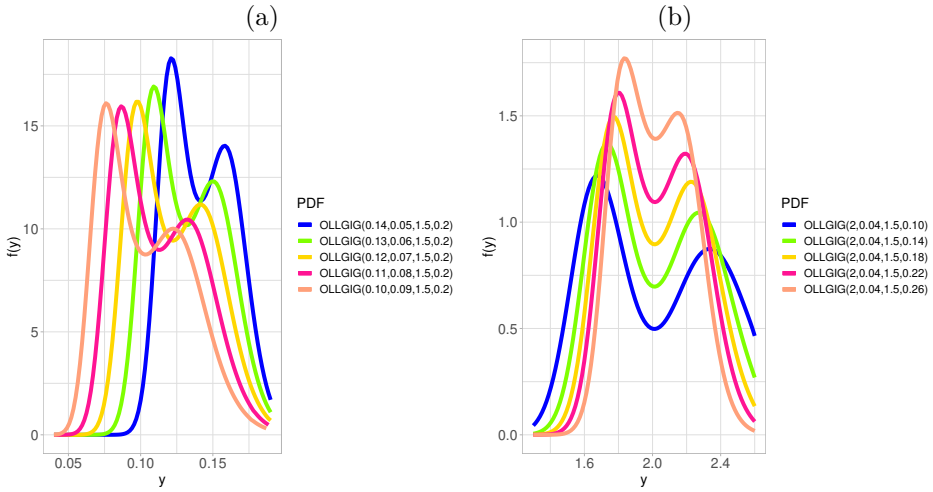

Figure 4.2. Plots of the OLLGIG density.

where $Q_{GIG}(u) = G^{-1}_{\mu,\sigma,\nu}(u)$ is the qf of the GIG distribution and $u \sim \text{Uniform}(0,1)$.

Consider a sample divided into N groups and Y_{ij} (for the j-th individual in the i-th group, $i = 1,\ldots,N$ and $j = 1,\ldots,n_i$) be independent random variables having the OLLGIG distribution. Each group has random effects W_i represented by independent and identically distributed random variables with density $g(w_i;\mathbf{V})$ and variance σ_w^2, where $\mathbf{V}$ is a vector of unknown parameters. By assuming that the random effects are unobserved random variables, the regression for correlated data can be expressed as

$$\mu_{ij} = \exp(\mathbf{x}_{ij}^T\boldsymbol{\beta} + w_i), \tag{4.6}$$

where $\mathbf{x}_{ij}^T = (x_{ij1},\ldots,x_{ijp})$ is the $p \times 1$ vector of covariates, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^T$ is the vector of unknown parameters, and w_i represents the random effects associated with the i-th group.

Further, assume that $Cov(W_i, Y_{ij}) = 0$ and that, conditioned on the random effects W_i, the response variables within the group i are independent with variance σ_w^2. So, the regression can be reduced to the classical regression when $\sigma_w^2 = 0$. For the random effect regression (4.6), the following assumptions hold:

- $y_{ij}|w_i \sim \text{OLLGIG}(\mu_{ij},\sigma,\nu,\tau)$, and marginal pdf

$$\begin{aligned} f(y_{ij}|w_i) &= \left(\frac{b}{\mu_{ij}}\right)^{\nu}\frac{\tau\, y_{ij}^{\nu-1}}{2K_\nu(\sigma^{-2})}\exp\left[-\frac{1}{2\sigma^2}\left(\frac{b y_{ij}}{\mu_{ij}}+\frac{\mu_{ij}}{b\, y_{ij}}\right)\right] \\ &\quad\times\{\eta(y_{ij})[1-\eta(y_{ij})]\}^{\tau-1}\{\eta(y_{ij})^\tau+[1-\eta(y_{ij})]^\tau\}^{-2}, \end{aligned} \tag{4.7}$$

  where

$$\eta(y_{ij}) = \int_0^{y_{ij}}\left(\frac{b}{\mu}\right)^{\nu}\frac{t^{\nu-1}}{2K_\nu(\sigma^{-2})}\exp\left[-\frac{1}{2\sigma^2}\left(\frac{b\,t}{\mu}+\frac{\mu}{b\,t}\right)\right]dt.$$

- The random variables $W_i \sim \text{N}(0,\sigma_w^2)$ (for $i = 1,\ldots,N$) have density

$$g(w_i;\mathbf{V}) = \frac{1}{\sqrt{2\,\pi}\,\sigma_w}\exp\left(-\frac{w_i^2}{2\sigma_w^2}\right), w_i \in \mathbb{R}. \tag{4.8}$$

The variance of W_i is $Var(W_i) = \sigma_w^2$. In this case, the parameter vector is $\mathbf{V} = \sigma_w^2$.

4.3 Estimation, simulations and residuals

The estimates of the parameters of the random effect OLLGIG regression are calculated via maximum likelihood. For each group i, the vector of the response variable is represented by $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{in_i})^T$. The likelihood function conditional on the random effects (independence within the group) for the individuals of the i-th group is

$$L_i(y_{ij}|w_i) = \prod_{j=1}^{n_i} f(y_{ij}|w_i), \tag{4.9}$$

where $f(y_{ij}|w_i)$ is the density (4.7). By assuming that the terms W_j and Y_{ij} are independent random variables, the contribution of the i-th group to the marginal likelihood function is

$$\int L_i(y_{ij}|w_i)\, g(w_i; \sigma w_i) dw_i,$$

where $g(\cdot)$ is the random effect density (4.8) and $L_i(y_{ij}|w_i)$ is given by (4.9).

Hence, under the assumption of independence between the N groups, the marginal likelihood function for the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma, \nu, \tau, \sigma_w)^\top$ reduces to

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f(y_{ij}|w_i)\, g(w_i, \sigma_w) dw_i. \tag{4.10}$$

Let $(y_{11}, \mathbf{x}_{11}), \ldots, (y_{1n_1}, \mathbf{x}_{1n_1}), \ldots, (y_{N1}, \mathbf{x}_{N1}), \ldots, (y_{Nn_N}, \mathbf{x}_{Nn_N})$ be $n = n_1 + \ldots + n_i$ observations, where y_{ij} is the response variable and $\mathbf{x}_{ij}$ is the vector of covariates associated with the j-th observation of the i-th group. Then, assuming the normal distribution (4.8) for the random effects and that Y is a random variable having the OLLGIG density (4.7), the logarithm of the marginal likelihood function (4.10) can be expressed as

$$\begin{aligned} l(\boldsymbol{\theta}) \;=\; & \sum_{i=1}^{N} \log \left\{ \int_{-\infty}^{+\infty} \prod_{j=1}^{n_i} \left(\frac{b}{\mu_{ij}}\right)^{\nu} \frac{\tau\, y_{ij}^{\nu-1}}{2K_\nu\,(\sigma^{-2})} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{b y_{ij}}{\mu_{ij}} + \frac{\mu_{ij}}{b\, y_{ij}}\right)\right] \times \right. \\ & \left. \prod_{j=1}^{n_i} \frac{\{\eta(y_{ij})[1-\eta(y_{ij})]\}^{\tau-1}}{\{\eta(y_{ij})^\tau + [1-\eta(y_{ij})]^\tau\}^2 \sqrt{2\pi}\sigma_w} \exp\left(-\frac{1}{2}\frac{w_i^2}{\sigma_w^2}\right) dw_i \right\}. \end{aligned} \tag{4.11}$$

The MLE $\widehat{\boldsymbol{\theta}}$ of the vector of parameters can be calculated by maximizing the log-likelihood (4.11) using the GAMLSS package in **R** software (Stasinopoulos et al., 2007). Initial values for $\boldsymbol{\beta}$, σ and σ_w can be taken from the fit of the IG regression with $\nu = 1$ and $\tau = 1$.

4.3.1 Simulation study

The quality of the MLEs of the parameters for the random effect OLLGIG regression is investigated via Monte Carlo simulations. One thousand replicates were performed for two groups ($N = 10$ and $N = 20$) with different sizes ($n_i = 5$, 25 and 70, $i = 1, \ldots, N$). A sample size n_i was generated for each replication from the OLLGIG$(\mu_{ij}, \sigma, \nu, \tau)$ distribution with $\nu = 1.5$ fixed under the configurations: $\sigma = 0.3$ and $\tau = 0.6$. For the parameters in μ_{ij}, the following values were taken: $\beta_0 = 0.15$ and $\beta_1 = 0.6$, and for the variance component $\sigma_w = 0.2$ (for $N = 10$) and $\sigma_w = 0.5$ (for $N = 20$). So, the parameter μ_{ij} has the systematic component $\mu_{ij} = \exp[(\beta_0 + w_i) + \beta_1\, x_{ij1}]$.

The response variable, the random effects and the explanatory variable were generated as:

- $y_{ij} \sim$ OLLGIG$(\mu_{ij}, \sigma, \nu, \tau)$;
- $W_i \sim$ Normal$(0, \sigma_w^2)$;

- $x_{ij} \sim$ Bernoulli(0.5).

Based on the results of the two scenarios ($\sigma_w = 0.2$, $N = 10$) and ($\sigma_w = 0.5$, $N = 20$) given in Table 4.1, it is noted that the MSEs decrease when n increases (as expected).

Table 4.1. Results of the simulation study: Scenario 1: ($\sigma_w = 0.2$, $N = 10$). Scenario 2 ($\sigma_w = 0.5$, $N = 20$).

		$n_i = 5$			$n_i = 25$			$n_i = 70$		
	Parameter	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
Scenario 1	β_0	0.232	0.082	0.024	0.139	-0.011	0.004	0.156	0.006	0.002
	β_1	0.624	0.024	0.019	0.595	-0.005	0.003	0.596	-0.004	0.001
	σ	0.465	0.165	0.129	0.301	0.001	0.019	0.269	-0.031	0.004
	τ	0.964	0.364	0.536	0.619	0.019	0.095	0.548	-0.052	0.023
	σ_w	0.179	-0.021	0.008	0.202	0.002	0.001	0.192	-0.008	<0.001
Scenario 2	β_0	0.376	0.226	0.060	0.385	0.235	0.057	0.264	0.114	0.013
	β_1	0.600	<0.001	0.008	0.598	-0.002	0.002	0.600	<0.001	0.001
	σ	0.348	0.048	0.046	0.288	-0.012	0.003	0.278	-0.022	0.002
	τ	0.714	0.114	0.212	0.585	-0.015	0.023	0.561	-0.039	0.009
	σ_w	0.284	-0.216	0.049	0.597	0.097	0.010	0.556	0.056	0.003

4.3.2 Residual analysis

For the new random effect regression, the quantile residuals (qrs) have the form

$$\widehat{qr}_{ij} = \Phi^{-1}\left\{\frac{\hat{\eta}^{\hat{\tau}}(y_{ij})}{\hat{\eta}^{\hat{\tau}}(y_{ij}) + [1-\hat{\eta}(y_{ij})]^{\hat{\tau}}}\right\}, \tag{4.12}$$

where

$$\eta(y_{ij}) = \int_0^{y_{ij}} \left(\frac{\hat{b}}{\hat{\mu}_{ij}}\right)^{\hat{\nu}} \frac{t^{\hat{\nu}-1}}{2\hat{K}_\nu(\hat{\sigma}^{-2})} \exp\left[-\frac{1}{2\hat{\sigma}^2}\left(\frac{\hat{b}\,t}{\hat{\mu}_{ij}} + \frac{\hat{\mu}_{ij}}{\hat{b}\,t}\right)\right] dt,$$

$$\hat{b} = \hat{K}_{\nu+1}(\hat{\sigma}^{-2})/\hat{K}_\nu(\hat{\sigma}^{-2}) \qquad \text{and} \qquad \hat{K}_\nu(t) = \frac{1}{2}\int_0^\infty y^{\hat{\nu}-1} \exp\left[-\frac{1}{2}t\left(u + u^{-1}\right)\right] du,$$

and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cdf.

Envelopes can be constructed from these residuals to provide better interpretation of the probability normal plots. The majority of the residuals will be randomly distributed within these bands if the regression is well-fitted.

Simulation study for the residuals

A simulation study is conducted to investigate the behavior of the empirical distribution of the residuals in (4.12). On thousand samples are generated via the algorithm described in Section 4.3.1. The normal probability plots are obtained for testing the normality of the residuals.

The residuals $\widehat{qr}_{ij}$ in (4.12) are calculated for each fitted regression. Figures 4.3 and 4.4 display the plots of the ordered residuals versus the expected values of the normal order statistics. These plots reveal that the empirical distribution of the qrs agrees with the standard normal distribution when n increases.

4.4 Application: Hectare price data

An application is now provided using the GAMLSS package in the **R** software (Stasinopoulos et al., 2007) to explain the average price per hectare of bare land in ten cities in the state of São Paulo (Brazil). These data come from the website of the Institute of Agricultural Economics (IEA) and the Office to Coordinate Integral Technical Assistance (CATI)[1] and refer to the two halves of 2015. The

[1] See: link: `http://ciagri.iea.sp.gov.br/bancoiea_TEste/Precor_TerraNua_SEFAZ.aspx`

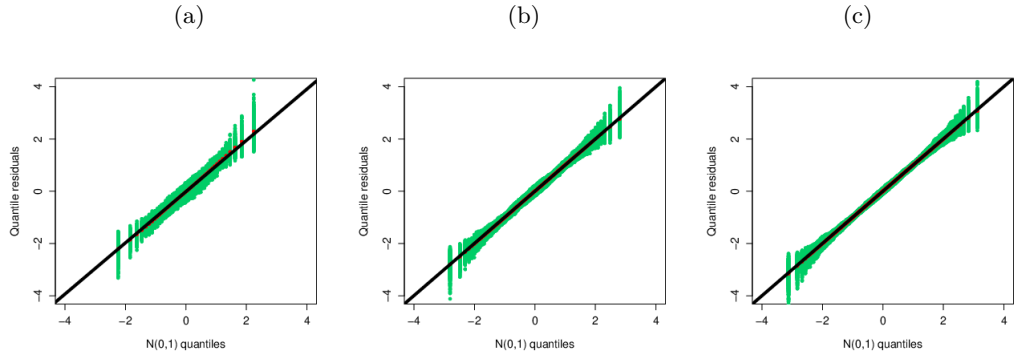

Figure 4.3. Normal probability plots for the qrs ($N = 10$) with $n = 5$, $n = 25$ and $n = 70$.

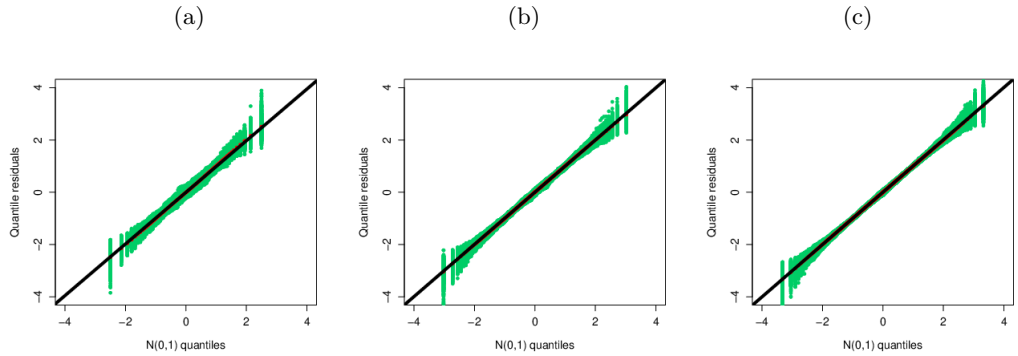

Figure 4.4. Normal probability plots for the qrs ($N = 20$) with $n = 5$, $n = 25$ and $n = 70$.

bare land price is defined as the commercial value of the land after deducting the value of structures, installations and other improvements, such as: buildings, storage sheds, barns and worker housing; stables, corrals, water pipes/hoses, aviaries, pigpens and other installations for shelter or management of animals; yards and similar areas for drying of agricultural products; rural electrification equipment; groundwater catchment and other installations for supply or distribution of water, including dams and tanks; fences; other improvements not related to rural activity; as well as perennial and temporary crops, cultivated and improved natural pastures; and planted forests.

The random effect OLLGIG regression is utilized to explain how the land categories of ten cities (Sorocaba, Adamantina, Águas de Lindóia, Alto Alegre, Bariri, Itapetininga, Itapeva, Santo André, São Carlos and Campinas) influence the average land price per hectare. A brief description of each city is now described. Sorocaba is one of the best cities in Brazil for investment in new start-ups as well as to live. It is near the capital, São Paulo. Adamantina has an area of approximately 411 square kilometers and is known for farming and stock breeding, and rural life in general. Águas de Lindóia is the capital of Brazil regarding hot springs, the reason why it is a cornerstone of the "Paulista Waters Circuit". Alto Alegre was served by Companhia Telefônica Brasileira (CTB) until 1973, and after that it was absorved into the Telecomunicações de São Paulo (TELESP), which constructed a central switching building that is still used today. In 1998, the company was sold to Telefônica as part of the privatization program, and in 2012 the company adopted the Vivo brand for fixed and cellular telephone operations. Bariri has a mixed industrial and agricultural base, in the latter case mostly sugarcane growing. Itapetininga is a large producer of corn, soybeans, oranges, milk and beef, as well as resins. Itapeva is an important producer of ores, especially phyllite, and also is among the leading municipalities in the state in the production of grain crops, besides having extensive reforested areas. Santo André has predominantly Atlantic Forest

vegetation, mainly in parks and environmental preservation areas. São Carlos is an important regional industrial center. Finally, Campinas is the state's largest city other than the capital, with a strong base of high-tech companies and educational institutions.

For this study, the variables are:

- y_{ij}: average price (R$) of a hectare of bare land (this variable was divided by 1,000);
- x_{ij1}: land categories (field land, primary cropland, secondary cropland, pasture land, reforestation land) (for $j = 1, \ldots, n_i,\ i = 1, \ldots, 10$).

Table 4.2 lists the averages and standard deviations (SDs) of the prices per hectare of bare land for each land category. The maximum price refers to the primary cropland, whereas the minimum price refers to the field land. The histogram of the average price per hectare (y_{ij}) in Figure 1 (Section 1) shows the presence of bimodality. So, for the marginal analysis, the OLLGIG distribution is capable to model these data.

Table 4.2. Averages and SDs for hectare price data.

land category	Average	SD
Field land	16.846	7.691
Primary cropland	29.971	11.829
Secondary cropland	25.656	10.803
Pasture land	22.810	9.003
Reforestation land	18.754	7.689

Table 4.3. Results from the fitted densities.

Distribution	$\log(\mu)$	$\log(\sigma)$	ν	τ
OLLGIG	21.132	3.700	23.118	0.324
	(0.936)	(0.931)	(13.348)	(0.144)
GIG	22.806	0.632	3.374	1
	(1.091)	(0.249)	(2.142)	(-)
IG	22.807	0.109	-0.5	1
	(1.192)	(0.008)	(-)	(-)

Table 4.3 gives the MLEs of the parameters and their standard errors (SEs) (in parentheses) from the fitted OLLGIG, GIG and IG distributions to the hectare prices. Table 4.4 lists the values of AIC (Akaike Information Criterion), CAIC (Consistent Akaike Information Criterion), BIC (Bayesian Information Criterion), HQIC (Hannan-Quinn information criterion), A^* (Anderson–Darling), W^* (Cramér-von Misses) and KS (Kolmogarov–Smirnov) for the fitted distributions. The results reveal that the OLLGIG distribution has the lowest values for these statistics, among the three. So, it could be chosen as the best distribution to explain the current data. Likelihood ration (LR) tests for comparing

Table 4.4. Some statistical measures.

Distribution	AIC	CAIC	BIC	HQIC	A^*	W^*	KS
OLLGIG	**744.634**	**745.055**	**755.054**	**748.851**	**0.238**	**0.031**	**0.050**
GIG	748.182	748.432	755.998	751.345	0.554	0.087	0.076
IG	749.077	749.201	754.288	751.186	0.863	0.138	0.093

the OLLGIG distribution with two special cases are reported in Table 4.5. The figures in this table (specially the p-values) reveal that the OLLGIG distribution provides a better fit to these data than the other two special cases.

Table 4.5. LR tests.

Models	Hypotheses	LR statistic	p-value
OLLGIG vs GIG	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	5.559	0.018
OLLGIG vs IG	$H_0 : \tau = 1, \nu = -0.5$ vs $H_1 : H_0$ is false	8.444	0.015

The empirical and estimated cdfs of the OLLGIG, GIG and IG distributions are given in Figure 4.5(a). The histogram of the data and the fitted OLLGIG, GIG and IG densities are displayed in Figure 4.5(b). These plots also reveal that the OLLGIG distribution provides the best fit to the hectare price data.

(a) (b)

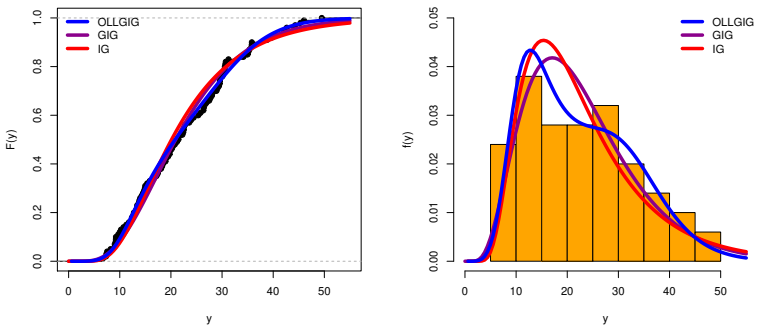

Figure 4.5. (a) Three estimated cdfs and empirical cdf. (b) Three estimated densities.

Regression analysis with systematic components

Four dummy variables $d_{ij1}, \ldots, d_{ij4}$ are defined since the covariable x_{ij1} has five categories of soil. According to the previous analysis, the random effect OLLGIG regression has the form (for $j = 1, \ldots, n_i,\ i = 1, \ldots, 10$)

$$\mu_{ij} = \exp\left(\beta_0 + \beta_1\, d_{ij1} + \beta_2\, d_{ij2} + \beta_3\, d_{ij3} + \beta_4\, d_{ij4} + w_i\right).$$

The random effect OLLGIG, GIG and IG regressions are compared via the AIC, BIC and GD (Global Deviance) statistics in Table 4.6. The wider regression outperforms the GIG and IG regressions irrespective of the criteria and then it can be used effectively in the analysis of these data.

Table 4.6. Statistics.

Model	AIC	BIC	GD
OLLGIG	**524.969**	**568.186**	**491.790**
GIG	528.885	570.248	497.131
IG	552.789	591.419	523.134

The results from the random effect OLLGIG regression fitted to these data via the GAMLSS package in **R** software are reported in Table 4.7. The explanatory variable x_{ij1} at the 5% level gives a significant difference among the levels of the land category to explain the price per hectare of bare land. The estimate of the variance component σ_w is also different from zero. So, it is necessary to consider random effects in the regression. Some interpretations are addressed at the end of this section.

Table 4.7. Results from the fitted random effect OLLGIG regression.

	Parameter	MLE	SE	p-value
Intercept	β_0	2.902	0.026	<0.001
primary cropland	β_1	0.608	0.040	<0.001
secondary cropland	β_2	0.441	0.042	<0.001
pasture land	β_3	0.322	0.043	<0.001
reforestation land	β_4	0.121	0.041	0.004
	$\log(\sigma)$	-0.697	0.076	
	ν	-4.167	0.961	
	τ	3.358	0.245	
	σ_w	0.424		

The random effect OLLGIG regression is compared with two special regressions via LR statistics in Table 4.8. The figures in this table indicate that the wider regression provides a better fit to these data than the other two regressions.

Table 4.8. LR tests.

Regressions	Hypotheses	LR statistic	p-value
OLLGIG vs GIG	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	5.341	0.012
OLLGIG vs IG	$H_0 : \tau = 1$ and $\nu = -0.5$ vs $H_1 : H_0$ is false	31.344	<0.001

Further, the qrs are plotted in Figure 4.6(a). These residuals are randomized around zero, thus revealing the suitability of the regression for analyzing the hectare price data. Finally, the quality of the adjustment range of the wider regression is verified by constructing the normal probability plot for the qrs with the simulated envelope. It is clear a good fitted regression shown in this envelope.

(a) (b)

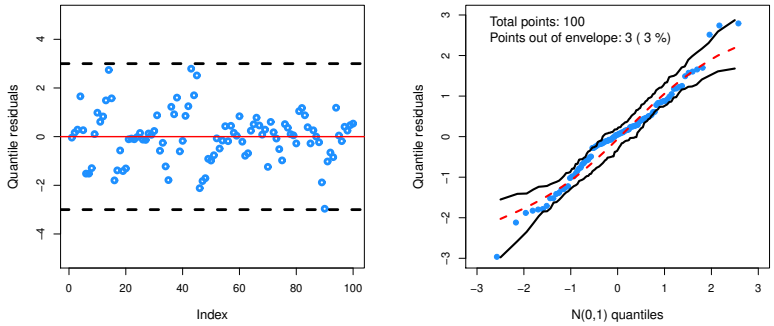

Figure 4.6. (a) Residual analysis of the random effect OLLGIG regression fitted to the hectare price data. (b) Normal probability plot for the qrs with envelope.

Table 4.9 provides the hypothesis testing to compare all levels of the covariate x_{ij1} considering the complete hectare price data. All levels of the land categories are significant (5%), and then there is strong evidence of differences between the price per hectare of bare land for each category.

The predicted random effects of the fitted regression are given in Table 4.10.

The hypothesis testing comparing the levels of the land category for each city separately with the associated predicted random effects are reported in the Appendix. Considering the 5% significance level, the interpretations based on these tables are addressed below.

Table 4.9. Hypothesis testing for bare land category levels.

Hypotheses H_0	Estimate	SE	p-value
primary cropland - field land = 0	0.608	0.040	<0.001
secondary cropland - field land = 0	0.441	0.042	<0.001
pasture land - field land = 0	0.322	0.043	<0.001
reforestation land - field land = 0	0.121	0.041	0.004
secondary cropland - primary cropland = 0	-0.167	0.045	<0.001
pasture land - primary cropland = 0	-0.286	0.045	<0.001
reforestation land - primary cropland = 0	-0.487	0.044	<0.001
pasture land - secondary cropland = 0	-0.119	0.047	0.013
reforestation land - secondary cropland = 0	-0.321	0.045	<0.001
reforestation land - pasture land = 0	-0.202	0.046	<0.001

Table 4.10. Prediction of the random effects.

w_i	Predicted	Cities
w_1	0.618	Sorocaba
w_2	-0.833	Adamantina
w_3	0.369	Águas de Lindóia
w_4	-0.331	Alto Alegre
w_5	-0.024	Bariri
w_6	0.139	Itapetininga
w_7	-0.339	Itapeva
w_8	-0.082	Santo André
w_9	-0.027	São Carlos
w_{10}	0.508	Campinas

- Table 4.11 gives the results of the comparisons between the categories of land in the city of Sorocaba in relation to the hectare price; there is no significant difference between the categories: secondary cropland - primary cropland.
- Table 4.12 shows that there is a difference between the categories of land: primary cropland - field land, pasture land - primary cropland and reforestation land - primary cropland in the Adamantina city in relation to the hectare price.
- Table 4.13 shows that there is a significant difference between all categories of land in the Águas de Lindóia city in relation to the hectare price.
- Table 4.14 shows that there is no significant difference between the categories of land in the Alto Alegre city in relation to the hectare price.
- Table 4.15 shows that there is a difference between the categories: primary cropland - field land, secondary cropland - field land, pasture land - field land, reforestation land - field land, reforestation land - primary cropland and reforestation land - secondary cropland in the Bariri city in relation to the hectare price.
- Table 4.16 shows that there is a significant difference between all categories of land in the Itapeniniga city in relation to the hectare price.
- Table 4.17 shows that there is a difference between the categories: primary cropland - field land, secondary cropland - field land, pasture land - field land, secondary cropland - primary croplands and pasture land - primary cropland in the Itapeva city in relation to the hectare price.
- Table 4.18 shows that there is a difference between the categories: primary cropland - field land, secondary cropland - field land, pasture land - field land, reforestation land - primary cropland,

reforestation land - secondary cropland and reforestation land - pasture land in the Santo André city in relation to the hectare price.

- Table 4.19 shows that there is no significant difference between the categories of land in the São Carlos city in relation to the hectare price.
- Table 4.20 shows the results of the comparisons between the categories of land in the city of Campinas in relation to the hectare price, note that there is no significant difference between the categories: reforestation land - field land secondary cropland - primary cropland reforestation land - pasture land.

4.5 Conclusions

This work presents a new random effect regression based on the *odd log-logistic Generalized inverse Gaussian* distribution. The new regression is a useful extension of some regressions with random effects and it can be a valuable option for data analysis when the response variable is bimodal in the areas of economics and agriculture. For different sample sizes, a simulation study was carried out to verify the consistency of the maximum likelihood estimates of the parameters. Quantile residuals are defined for the proposed regression. The versatility of the new regression is proved for estimating the average hectare value of raw land.

Appendix

Table 4.11. Hypothesis testing for land category levels to Sorocaba city.

primary cropland - field land = 0	0.467	0.031	<0.001
secondary cropland - field land = 0	0.384	0.032	<0.001
pasture land - field land = 0	0.221	0.029	0.001
reforestation land - field land = 0	0.031	0.052	0.578
secondary cropland - primary cropland = 0	-0.084	0.028	0.032
pasture land - primary cropland = 0	-0.247	0.027	<0.001
reforestation land - primary cropland = 0	-0.436	0.050	<0.001
pasture land - secondary cropland = 0	-0.163	0.028	0.002
reforestation land - secondary cropland = 0	-0.353	0.051	<0.001
reforestation land - pasture land = 0	-0.189	0.049	0.013

Table 4.12. Hypothesis testing for land category levels to Adamantina city .

primary cropland - field land = 0	0.425	0.137	0.027
secondary cropland - field land = 0	0.271	0.135	0.101
pasture land - field land = 0	0.176	0.136	0.252
reforestation land - field land = 0	0.163	0.135	0.280
secondary cropland - primary cropland = 0	-0.154	0.091	0.149
pasture land - primary cropland = 0	-0.249	0.093	0.044
reforestation land - primary cropland = 0	-0.262	0.091	0.035
pasture land - secondary cropland = 0	-0.094	0.089	0.339
reforestation land - secondary cropland = 0	-0.108	0.087	0.272
reforestation land - pasture land = 0	-0.013	0.089	0.889

Table 4.13. Hypothesis testing for land category levels to Águas de Lindóia city.

primary cropland - field land = 0	0.759	0.059	<0.001
secondary cropland - field land = 0	0.555	0.033	<0.001
pasture land - field land = 0	0.404	0.034	<0.001
reforestation land - field land = 0	0.171	0.018	<0.001
secondary cropland - primary cropland = 0	-0.202	0.066	0.028
pasture land - primary cropland = 0	-0.354	0.067	0.003
reforestation land - primary cropland = 0	-0.586	0.060	<0.001
pasture land - secondary cropland = 0	-0.151	0.044	0.018
reforestation land - secondary cropland = 0	-0.384	0.033	<0.001
reforestation land - pasture land = 0	-0.232	0.034	0.001

Table 4.14. Hypothesis testing for land category levels to Alto Alegre city.

Hypotheses H_0	Estimate	SE	p-value
primary cropland - field land = 0	0.418	0.192	0.081
secondary cropland - field land = 0	0.299	0.204	0.203
pasture land - field land = 0	0.284	0.209	0.232
reforestation land - field land = 0	0.025	0.171	0.891
secondary cropland - primary cropland = 0	-0.119	0.237	0.636
pasture land - primary cropland = 0	-0.133	0.241	0.604
reforestation land - primary cropland = 0	-0.393	0.209	0.119
pasture land - secondary cropland = 0	-0.014	0.251	0.957
reforestation land - secondary cropland = 0	-0.274	0.220	0.269
reforestation land - pasture land = 0	-0.259	0.225	0.301

Table 4.15. Hypothesis testing for land category levels to Bariri city.

primary cropland - field land = 0	0.954	0.176	0.003
secondary cropland - field land = 0	0.778	0.131	0.002
pasture land - field land = 0	0.623	0.139	0.007
reforestation land - field land = 0	0.268	0.094	0.036
secondary cropland - primary cropland = 0	-0.177	0.201	0.419
pasture land - primary cropland = 0	-0.331	0.207	0.169
reforestation land - primary cropland = 0	-0.687	0.179	0.012
pasture land - secondary cropland = 0	-0.155	0.170	0.405
reforestation land - secondary cropland = 0	-0.509	0.136	0.013
reforestation land - pasture land = 0	-0.355	0.144	0.057

Table 4.16. Hypothesis testing for land category levels to Itapetininga city.

primary cropland - field land = 0	0.572	0.027	<0.001
secondary cropland - field land = 0	0.404	0.009	<0.001
pasture land - field land = 0	0.301	0.006	<0.001
reforestation land - field land = 0	0.101	0.008	<0.001
secondary cropland - primary cropland = 0	-0.167	0.028	0.002
pasture land - primary cropland = 0	-0.271	0.027	<0.001
reforestation land - primary cropland = 0	-0.471	0.027	<0.001
pasture land - secondary cropland = 0	-0.104	0.008	<0.001
reforestation land - secondary cropland = 0	-0.304	0.009	<0.001
reforestation land - pasture land = 0	-0.199	0.013	0.004

Table 4.17. Hypothesis testing for land category levels to Itapeva city.

primary cropland - field land = 0	0.699	0.061	<0.001
secondary cropland - field land = 0	0.474	0.056	<0.001
pasture land - field land = 0	0.339	0.071	0.005
reforestation land - field land = 0	-0.157	0.280	0.599
secondary cropland - primary cropland = 0	-0.225	0.065	0.018
pasture land - primary cropland = 0	-0.359	0.079	0.006
reforestation land - primary cropland = 0	-0.489	0.909	0.614
pasture land - secondary cropland = 0	-0.134	0.074	0.127
reforestation land - secondary cropland = 0	-0.614	0.268	0.071
reforestation land - pasture land = 0	-0.094	0.904	0.921

Table 4.18. Hypothesis testing for land category levels to Santo André city.

primary cropland - field land = 0	0.889	0.182	0.004
secondary cropland - field land = 0	0.598	0.076	0.001
pasture land - field land = 0	0.548	0.070	0.001
reforestation land - field land = 0	0.141	0.121	0.298
secondary cropland - primary cropland = 0	-0.291	0.183	0.172
pasture land - primary cropland = 0	-0.341	0.181	0.117
reforestation land - primary cropland = 0	-0.749	0.206	0.015
pasture land - secondary cropland = 0	-0.049	0.073	0.527
reforestation land - secondary cropland = 0	-0.457	0.123	0.014
reforestation land - pasture land = 0	-0.408	0.119	0.019

Table 4.19. Hypothesis testing for land category levels to São Carlos city.

primary cropland - field land = 0	0.393	0.253	0.182
secondary cropland - field land = 0	0.278	0.251	0.318
pasture land - field land = 0	0.199	0.239	0.443
reforestation land - field land = 0	0.147	0.224	0.540
secondary cropland - primary cropland = 0	-0.115	0.272	0.691
pasture land - primary cropland = 0	-0.194	0.261	0.491
reforestation land - primary cropland = 0	-0.245	0.248	0.367
pasture land - secondary cropland = 0	-0.079	0.259	0.772
reforestation land - secondary cropland = 0	-0.131	0.245	0.617
reforestation land - pasture land = 0	-0.052	0.233	0.834

Table 4.20. Hypothesis testing for land category levels to Campinas city.

primary cropland - field land = 0	0.418	0.057	<0.001
secondary cropland - field land = 0	0.312	0.054	0.002
pasture land - field land = 0	0.171	0.053	0.024
reforestation land - field land = 0	0.040	0.079	0.629
secondary cropland - primary cropland = 0	-0.106	0.050	0.088
pasture land - primary cropland = 0	-0.247	0.049	0.004
reforestation land - primary cropland = 0	-0.378	0.076	0.004
pasture land - secondary cropland = 0	-0.141	0.046	0.028
reforestation land - secondary cropland = 0	-0.272	0.074	0.014
reforestation land - pasture land = 0	-0.131	0.074	0.136

References

Coupé, C. (2018). Modeling linguistic variables with regression models: Addressing non-gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized

additive models for location, scale, and shape. *Frontiers in Psychology*, 9:513.

Crowder, M. J. and Hand, D. J. (1990). *Analysis of repeated measures*, volume 41. CRC Press.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, pages 959–971.

Dirmeier, S., Dächert, C., van Hemert, M., Tas, A., Ogando, N. S., van Kuppeveld, F., Bartenschlager, R., Kaderali, L., Binder, M., and Beerenwinkel, N. (2020). Host factor prioritization for pan-viral genetic perturbation screens using random intercept models and network propagation. *PLoS Computational Biology*, 16(2):e1007587.

Hashimoto, E. M., Silva, G. O., Ortega, E. M., and Cordeiro, G. M. (2019). Log-burr xii gamma–weibull regression model with random effects and censored data. *Journal of Statistical Theory and Practice*, 13(2):1–21.

Ho, N. T., Li, F., Wang, S., and Kuhn, L. (2019). metamicrobiomer: an r package for analysis of microbiome relative abundance data using zero-inflated beta gamlss and meta-analysis across studies using random effects models. *BMC Bioinformatics*, 20(1):188.

Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution*. Springer Science & Business Media, New York.

Muniz-Terrera, G., Hout, A. v. d., Rigby, R., and Stasinopoulos, D. (2016). Analysing cognitive test data: Distributions and non-parametric random effects. *Statistical Methods in Medical Research*, 25(2):741–753.

Souza Vasconcelos, J. C., Cordeiro, G. M., Ortega, E. M., and Araújo, E. G. (2019). The new odd log-logistic generalized inverse gaussian regression model. *Journal of Probability and Statistics*, 2019.

Stasinopoulos, D. M., Rigby, R. A., et al. (2007). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46.

5 FINAL CONSIDERATIONS

In this thesis, I develop flexible parametric and semiparametric regression models based on an extension of the generalized inverse Gaussian distribution and on the odd log-logistic generator of distributions. The new proposed distribution is called the odd log-logistic generalized inverse Gaussian (OLLGIG). The main characteristic of this distribution is that it permits modeling bimodal data without the need to use a mixture of distributions. I use the gamlss package available in the **R** software to obtain the maximum likelihood and penalized maximum likelihood estimates (MLE and PMLE), as well as to evaluate the sensitivity (global influence and analysis of residuals) of the proposed regression models. In Chapter 2, I define the OLLGIG distribution and describe various structural properties. I then present various simulation studies to evaluate the performance of the MLEs and to study the distribution and residuals utilized empirically to validate the assumptions of the proposed regression models. Finally, I present two applications to real data, the first without considering covariables and the second considering covariables in two systematic components. In Chapter 3, I propose additive, partial and semiparametric regression models based on the OLLGIG distribution and consider three types of penalized smoothers, as well as discussing selection criteria, sensitivity analysis and residuals. Finally, data on climatology, ethanol use and air quality are used to verify the versatility of the proposed models. In Chapter 4 I propose the regression model with fixed effect in the intercept based on the OLLGIG distribution, applying it to a dataset on land values per hectare in some cities in the state of São Paulo, Brazil.