UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

NICOLLAS STEFAN SOARES DA COSTA

EXAMINING THE GENERALIZED ODD LOG-LOGISTIC FAMILY: A REGRESSION COMPILATION

Recife

2024

NICOLLAS STEFAN SOARES DA COSTA

EXAMINING THE GENERALIZED ODD LOG-LOGISTIC FAMILY: A REGRESSION COMPILATION

Tese apresentada ao programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco como requisito parcial para obtenção do título de Doutor em Estatística.

Área de Concentração: Probabilidade Estatística

Orientadora: Profa. Dra. Maria do Carmo Soares de Lima

Coorientador: Prof. PhD. Gauss Moutinho Cordeiro

Recife

2024

Catalogação na fonte
Bibliotecário Josias Machado da Silva Junior, CRB4-1690

C837e Costa, Nicollas Stefan Soares da.
Examining the generalized odd log-Logistic Family: a regression compilation / Nicollas Stefan Soares da Costa. – 2024.
131 f.: il., fig., tab.

Orientadora: Maria do Carmo Soares de Lima.
Coorientador: Gauss Moutinho Cordeiro
Tese (Doutorado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2024.
Inclui referências.

1. Diagnóstico. 2. Família generalizada odd log-logística. 3. Máxima Verossimilhança. 4. Modelo de regressão. 5. Simulação. I. Lima, Maria do Carmo Soares de (orientadora). II. Cordeiro, Gauss Moutinho (orientador). III. Título.

310 CDD (23. ed.) UFPE- CCEN 2024 - 34

NICOLLAS STEFAN SOARES DA COSTA

EXAMINING THE GENERALIZED ODD LOG-LOGISTIC FAMILY: A REGRESSION COMPILATION

Tese apresentada ao Programa de Pós Graduação em Estatística da Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, como requisito para a obtenção do título de Doutor em Estatística. Área de concentração: Estatística Aplicada.

Aprovado em: 03/04/2024.

BANCA EXAMINADORA

_____
Prof[a]. Dr[a]. Maria do Carmo Soares de Lima (Orientadora)
Universidade Federal de Pernambuco - UFPE

_____
Prof. PhD. Gauss Moutinho Cordeiro (Coorientador)
Universidade Federal de Pernambuco - UFPE

_____
Prof. Dr. Fernando Arturo Peña Ramírez (Examinador Externo)
Universidade do Contestado - UNC

_____
Prof. Dr. Josimar Mendes de Vasconcelos (Examinador Externo)
Universidade Federal Rural de Pernambuco - UFRPE

_____
Prof[a]. Dr[a]. Renata Rojas Guerra (Examinadora Externo)
Universidade Federal de Santa Maria - UFSM

_____
Prof. Dr. Abraão David Costa do Nascimento (Examinador Interno)
Universidade Federal de Pernambuco - UFPE

To my parents, **Efigênia Maria Soares da Costa** and **Murilo Vieira da Costa** (*in memoriam*), who have always been an example in my life.

"*Data! Data! Data!... I can't make bricks without clay.*" Sherlock Holmes, The Adventure of the Copper Beeches (Arthur Conan Doyle).

RESUMO

In this work, considering the family of distributions, generalized odd log-logistic-G, several applications have been proposed with different real data using regression models. The distributions of this family accommodate asymmetric, bimodal and heavy-tailed forms, showing flexibility when compared to other well-known generator distributions. Based on the generator family of distributions presented, regression models have been introduced with distinct systematic structures, linking the explanatory variables through the parameters of the baseline distribution and all computational modeling is implemented using the R software. The first two applications involve two univariate distributions: Lindley and exponential. The first uses the novel generalized odd log-logistic Lindley distribution to evaluate data on the completed primary vaccination rate of COVID-19 in counties in the American state of Texas. The second uses the generalized odd log-logistic exponential distribution to investigate dengue fever weekly cases in the Federal District of Brazil. The other applications relied on the well-known continuous distributions, gamma, and Weibull distributions. The first applies the generalized odd log-logistic gamma distribution to agricultural data on yacon potatoes from a study in Peru. The following analysis employs the generalized odd log-logistic Weibull distribution to examine daily wind power generation data in Brazil. Monte Carlo simulations are used to evaluate the accuracy of maximum likelihood estimates using a variety of measures. In order to determine the most suitable model, the research includes goodness-of-fit measures, diagnostics and residual analysis. Finally, the findings obtained utilizing various data sets demonstrated that the proposed models are a viable alternative to competing distributions.

Keywords: diagnostic; generalized odd log-logistic family; maximum likelihood; regression model; simulation.

ABSTRACT

Neste trabalho, considerando a família de distribuições log-logística odd generalizada-G, foram propostas várias aplicações com diferentes dados reais usando modelos de regressão. As distribuições dessa família acomodam formas assimétricas, bimodais e de cauda pesada, mostrando flexibilidade quando comparadas a outras distribuições de geradores conhecidos. Com base na classe geradora de distribuições apresentada, foram introduzidos modelos de regressão com estruturas sistemáticas distintas, vinculando as variáveis explicativas por meio dos parâmetros da distribuição *baseline* e toda a modelagem computacional foi implementada usando o *software* R. As duas primeiras aplicações envolvem duas distribuições univariadas: Lindley e exponencial. A primeira usa a nova distribuição Lindley log-logística *odd* generalizada para avaliar dados sobre a taxa de vacinação primária completa de COVID-19 em condados do estado Americano do Texas. A segunda usa a distribuição exponencial log-logística *odd* generalizada para investigar casos semanais de dengue no Distrito Federal do Brasil. As outras aplicações se basearam nas conhecidas distribuições contínuas, gama e Weibull. A primeira aplica a distribuição gama log-logística *odd* generalizada a dados agrícolas sobre batatas yacon de um estudo no Peru. A análise seguinte emprega a distribuição Weibull log-logística *odd* generalizada para examinar os dados diários de geração de energia eólica no Brasil. Simulações de Monte Carlo são utilizadas para avaliar a acurácia das estimativas de máxima verosimilhança utilizando uma variedade de medidas. Para determinar o modelo mais adequado, a investigação inclui medidas de adequação, diagnóstico e análise de resíduos. Por fim, as conclusões obtidas com o uso de vários conjuntos de dados demonstraram que os modelos propostos são uma alternativa viável às distribuições concorrentes.

Palavras-chaves: diagnóstico; família generalizada *odd* log-logística; máxima verossimilhança; modelo de regressão; simulação.

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

GOLL-G	Generalized Odd Log-Logistic-G
COVID-19	Corona Virus Disease 2019
US	United States
GOLLFW	Generalized Odd Log-Logistic Flexible Weibull
GOLLE	Generalized Odd Log-Logistic Exponential
GOLLMax	Generalized Odd Log-Logistic Maxwell
GOLLNH	Generalized Odd Log-Logistic Nadarajah Haguigui
GOLLBS	Generalized Odd Log-Logistic Birnbaum-Saunders
GOLLLN	Generalized Odd Log-Logistic Log-Normal
GOLLFr	Generalized Odd Log-Logistic Fréchet
GOLLGR	Generalized Odd Log-Logistic Generalized Rayleigh
GOLLBE	Generalized Odd Log-Logistic Beta
GOLLMaxMix	Generalized Odd Log-Logistic Maxwell Mixture
cdf	Cumulative Distribution Function
T-X	Transformer-Transformer Generator
OLL-G	Odd Log-Logistic-G
rv	Random Variable
GOLLN	Generalized Odd Log-Logistic Normal
pdf	Probability Density Function
hrf	Hazard Rate Function
Exp-G	Exponentiated-G
qf	Quantile Function

mgf	Moment Generation Function
MLE	Maximum Likelihood Estimate
AE	Average Estimate
MSE	Mean Square Error
AB	Absolute Bias
AL	Estimated Average Length
RMSE	Root Mean Square Error
CP	Coverage Probability
SE	Standard Error
GoF	Goodness-of-Fit
AIC	Akaike Information Criterion
CAIC	Consistent Akaike Information Criterion
BIC	Bayesian Information Criterion
W^*	Cramér-von Mises
A^*	Anderson-Darling
KS	Kolmogorov-Sminorv
LR	Likelihood Ratio Statistic
GCD	Generalized Cook Distance
LD	Loglikelihood Distance
EUA	Estados Unidos da América
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
GOLLL	Generalized Odd Log-Logistic Lindley
OLLL	Odd Log-Logistic Lindley

Exp-L	Exponentiated-Lindley
L	Lindley
W_{-1}	Lambert W Function
BL	Beta-Lindley
KwL	Kumaraswamy-Lindley
GL	Gamma-Lindley
WL	Weibull-Lindley
$I_{G(x)}(a,b)$	Incomplete Beta Function
BFGS	Broyden–Fletcher–Goldfarb–Shanno Algorithm
CDC	Centers for Disease Control and Prevention
VR	Population Rate with Complete Primary Series of COVID-19 Vaccination
HP	Hospitals Reporting vaccination
PR	Poverty Rate
MS	Metropolitan Status
HR	High School Completion Rate
BA	Broadband Acess
HT	Heart Disease Rate
GEV	Generalized Extreme Value
EVT	Extreme Value Theory
GOLLE	Generalized Odd Log-Logistic Exponential
OLLE	Odd Log-Logistic Exponential
Exp-E	Exponentiated-Exponential
E	Exponential

KwFr	Kumaraswamy-Fréchet
KwE	Kumaraswamy-Exponential
GFr	Gamma-Fréchet
GE	Gamma-Exponential
BE	Beta-Fréchet
Fr	Fréchet
SES-DF	Federal District Health Department
DG	Dengue Fever Cases
GOLLΓ	Generalized Odd Log-Logistic Gamma
OLLΓ	Odd Log-Logistic Gamma
Exp-Γ	Exponetiated-Gamma
Γ	Gamma
GG	Generalized Gamma
$G_{2,3}^{3,0}\left(.\,\middle|\,\cdot\right)$	Meijer G-Function
KwΓ	Kumaraswamy-Gamma
WΓ	Weibull-Gamma
NWΓ	New Weibull-Gamma
NFΓ	New Flexible Gamma
IPC	International Potato Center
DBR	Degree Brix
SUCR	Sucrose
LOC1	Dummy Variable of Location 1
LOC2	Dummy Variable of Location 2

FOS	Fructo-Oligosaccharides
FRUC	Fructose
OLLExGa	Odd Log-Logistic Exponential Gaussian
GOLLW	Generalized Odd Log-Logistic Weibull
OLLW	Odd Log-Logistic Weibull
Exp-W	Exponentiated-Weibull
OLLR	Odd Log-Logistic Rayleigh
Exp-R	Exponentiated-Rayleigh
R	Rayleigh
BW	Beta-Weibull
KwW	Kumaraswamy-Weibull
GW	Gamma-Weibull
KwFr	Kumaraswamy-Fréchet
MW	Megawatt
ONS	Operador Nacional do Sistema Elétrico

CONTENTS

1 INTRODUCTION

Regression analysis is a valuable tool statistical technique with a wide application in many scientific fields. Traditionally, in linear regression, when the normality assumptions of the response variable are not satisfied, researchers apply some type of transformation. This approach is used because many phenomena have asymmetry, bimodal form or the presence of discrepant observations. Due to the limitations of this process, lately, the development of new regression models has grown to handle non-normal data sets. For instance, novel models can produce more accurate and insightful results.

Exploring new distributions provides more powerful methods for expanding regression's applicability, improving prediction accuracy, their structures and allowing for greater flexibility in fitting real data. This work focuses on the *generalized odd log-logistic-G* (GOLL-G) distribution, pioneered by Cordeiro et al. (2017). In light of this, a linear representation is presented, some mathematical features are addressed and the maximum likelihood of the class of GOLL-G distributions is described. The family of distribution's applications are illustrated using a regression structure on the baseline parameter vector using data on the corona virus disease 2019 (COVID-19) immunization rate in 254 Texas, United States (US) counties in a window scenario in February 2023, dengue fever weekly cases in the Federal District of Brazil in 2022, yacon potato sugar concentration in a Peruvian research study and the total daily Brazilian wind energy generation in 2022.

In recent years, the GOLL-G class has applications in several domains, demonstrating its superiority over well-known distributions. Prataviera et al. (2018) introduced the GOLL-flexible Weibull (GOLLFW) distribution to modeling repairable systems data (sugarcane harvester machines). One notable feature of this new regression model is that it does not require the assumption of proportional risks. The generalized odd log-logistic exponential (GOLLE) distribution (QOSHJA; MUÇA, 2018) is applied in fatigue fracture of Kevlar 373/epoxy data and confirms the better fit than beta and Kumaraswamy generators. Moreover, the GOLL-Maxwell (GOLLMax) distribution (PRATAVIERA; ORTEGA; CORDEIRO, 2020) is presented to study bimodal data in three engineering applications (strength, image and brittle materials data), whose key advantage over other competing distributions is modelling bimodal, asymmetric and heavy tail data. In addition, Prataviera et al. (2020) proposed a novel GOLLMax regression with application to microbiology data. The major characteristic is that it fits data

with covariables in the presence of bimodality, heteroscedasticity, zero-inflation and nonlinear effects.

Additionally, Ibrahim (2020) introduced a new four-parameter distribution named generalized odd log-logistic Nadarajah Haghighi (GOLLNH), studied several mathematical properties and used different methods of estimation. The flexibility of the new model is illustrated by the many important shapes of the failure rate function and some applications. Afify et al. (2021) defined the GOLLE distribution for application in reparable systems. The model can model various shapes of aging and failure criteria. Further, the estimation difficulties of unknown parameters under complete and type II censored data were also examined, utilizing the maximum likelihood and Bayesian estimation approaches. Furthermore, Vasconcelos et al. (2022) proposed a parametric and a partially linear regression based on the generalized odd log-logistic Birnbaum–Saunders (GOLLBS) distribution using average price data received by producers and wholesalers. Compared to competitive regression models in the literature, the new model's flexibility has a better fit. Likewise, Vasconcelos et al. (2022) presented a parametric and additive partial linear regressions based on the generalized odd log-logistic log-normal (GOLLLN) distribution to employ in a climatological and an aerobic stability test data. The model proved essential for identifying covariables with nonlinear effects on the response variable.

Khaleq (2022) defined the generalized odd log-logistic Fréchet (GOLLFr) distribution for modeling extreme values. Three applications are addressed to breaking stress, glass fibers and relief time data and showed better fit to other distributions. Cordeiro et al. (2022) proposed the generalized odd log-logistic generalized Rayleigh (GOLLGR) distribution. The researchers constructed a novel regression using the lifetimes of patients diagnosed with COVID-19 and examined the effect of age and diabetes on the time until death. Both covariates are found to be significant in the response variable. Moreover, Prataviera et al. (2022) model rates and proportions data with the new generalized odd log-logistic beta (GOLLBE) distribution. Lastly, Vigas et al. (2023) used the GOLL-G family to propose a new regression for interval-censored survival data and Prataviera et al. (2023) described the GOLLMax mixture (GOLLMaxMix) model to analyze COVID-19 Chinese data.

In the context of the GOLL-G family of distributions presented, the thesis is organized in the following manner and each Chapter can be read independently. Chapter 2 reviews the class of distributions called generalized odd log-logistic-G (CORDEIRO et al., 2017), presents some of these properties and uses the maximum likelihood approach to estimate the parameters.

An innovative regression model based on the GOLL-G distribution is proposed, linking the covariates into the vector parameter of the baseline distribution and simulations are computed using some measures to evaluate the accuracy of the MLEs. The best fit model is evaluated using goodness-of-fit measurements, diagnostics and residual analysis. All computational requirements of the new distribution and regression model were implemented in the R software. Chapters 3 and 4 demonstrated the use of the GOLL-G with two univariate distributions: Lindley and exponential. The first one uses the new generalized odd log-logistic Lindley distribution with a shape regression framework to analyze counties Texas, US COVID-19 vaccination rate data. The second addresses by Qoshja e Muça (2018) and Afify et al. (2021) defined the generalized odd log-logistic exponential distribution with a regression structure in the shape parameter applied to dengue fever weekly cases data in Federal District, Brazil. Chapters 5 and 6 present two regression models based on the well-known continuous distributions, gamma and Weibull. The primary one presented by Cordeiro et al. (2017) is the generalized odd log-logistic gamma distribution with shape and scale regression structure for agricultural data. The next one used a novel generalized odd log-logistic Weibull distribution Cordeiro et al. (2017) with a bivariate regression structure for total daily Brazilian wind energy generation data. Finally, some considerations and perspectives are addressed in Chapter 7.

2 REVIEW OF THE GENERALIZED ODD LOG-LOGISTIC FAMILY OF DISTRIBUTIONS

RESUMO

A família de distribuições log-logística *odd* generalizada é apresentada como uma revisão e foi introduzida de forma pioneira por Cordeiro et al. (2017). Trata-se de uma classe de distribuições contínuas com dois parâmetros de forma extra que produzem formas flexíveis para ajustar dados reais. Uma representação linear da família é apresentada e algumas de suas propriedades matemáticas, incluindo a função quantílica, os momentos e a função geradora de momentos são abordadas. É definido o método de máxima verossimilhança para estimar os parâmetros e maximizá-los numericamente usando o R Core Team (2021). O modelo de regressão baseado na classe de distribuições é apresentado com a proposta de vincular as variáveis explicativas por meio do vetor de parâmetros da distribuição *baseline*. O comportamento das estimativas é estudado por meio de simulações de Monte Carlo usando algumas medidas. Para avaliar o modelo de melhor ajuste, são usadas algumas medidas de bondade de ajuste, diagnóstico e análise de resíduos.

Palavras-chaves: Diagnóstico. Distribuição base. Família log-logística *odd* generalizada. Máxima verossimilhança. Medidas de bondade. Propriedades. Regressão. Resíduos. Simulação.

ABSTRACT

The family of generalized odd log-logistic distributions is presented as a review and was pioneered introduced by Cordeiro et al. (2017). It is a class of continuous distributions with two extra shape parameters that produce very flexible forms for fitting real data. A linear representation of the family is presented and some of its mathematical properties, including the quantile function, the moments and the moment generating function are addressed. The method of maximum likelihood to estimate the parameters and maximize them numerically using the R Core Team (2021) is defined. The regression model based on the class of distributions is presented with the proposal to link the explanatory variables through the parameter vector of the baseline distribution. The behavior of the estimates is studied by means of Monte Carlo simulations using some measures. To assess the best fit model, some goodness-of-fit measures, diagnostics and residuals analyses are used.

Keywords: Baseline. Diagnostic. Generalized odd log-logistic family. Goodness-of-fit. Maximum likelihood. Properties. Regression. Residuals. Simulation.

2.1 INTRODUCTION

Consider a baseline cumulative distribution function (cdf), $G(x;\boldsymbol{\xi})$, with unknown parameter vector $\boldsymbol{\xi}$. The cdf of the odd log-logistic (OLL-G) family (GLEATON; LYNCH, 2006) is characterized, as:

$$F(y;\alpha,\boldsymbol{\xi}) = \frac{G(x;\boldsymbol{\xi})^{\alpha}}{G(x;\boldsymbol{\xi})^{\alpha}+[1-G(x;\boldsymbol{\xi})]^{\alpha}}, \tag{2.1}$$

where $\alpha > 0$.

Based on the transformer-transformer (T-X) generator (ALZAATREH; LEE; FAMOYE, 2013), Cordeiro et al. (2017) introduced the cdf of the *generalized odd log-logistic* (GOLL-G) family of distributions integrating the log-logistic density function, as follows:

$$F(y;\alpha,\theta,\boldsymbol{\xi}) = \int_{0}^{\frac{G(x;\boldsymbol{\xi})^{\theta}}{1-G(x;\boldsymbol{\xi})^{\theta}}} \frac{\alpha t^{\alpha-1}}{(1-t^{\alpha})^{2}}dt = \frac{G(x;\boldsymbol{\xi})^{\alpha\theta}}{G(x;\boldsymbol{\xi})^{\alpha\theta}+[1-G(x;\boldsymbol{\xi})^{\theta}]^{\alpha}}, \tag{2.2}$$

where $\alpha > 0$ and $\theta > 0$ are two extra shape parameters that play an important role in shaping the overall shape, skewness, tail behavior and central tendency of the distribution, allowing to model a variety of data sets.

Let $Y \sim$ GOLL-G$(\alpha,\theta,\boldsymbol{\xi})$ be a random variable (rv) with cdf (2.2). Differentiating Equation (2.2), the probability density function (pdf) of Y reduces to:

$$f(y;\alpha,\theta,\boldsymbol{\xi}) = \frac{\alpha\theta g(x;\boldsymbol{\xi})G(x;\boldsymbol{\xi})^{\alpha\theta-1}[1-G(x;\boldsymbol{\xi})^{\theta}]^{\alpha-1}}{\{G(x;\boldsymbol{\xi})^{\alpha\theta}+[1-G(x;\boldsymbol{\xi})^{\theta}]^{\alpha}\}^{2}}, \tag{2.3}$$

where $g(x;\boldsymbol{\xi})$ is the parent pdf. Due to its great flexibility, firstly shown through special models in Cordeiro et al. (2017), the density function (2.3) is widely used in many areas, such as biology (GOLLMax), engineering (GOLLFW, GOLLE and GOLLMax), economy (GOLLBS), epidemiology (GOLLGR, generalized odd log-logistic normal (GOLLN) and GOLLMaxMix), extreme value theory (GOLLFr), meteorology (GOLLLN), among others.

The hazard rate function (hrf) of Y is defined, as:

$$\tau(y;\alpha,\theta,\boldsymbol{\xi}) = \frac{\alpha\theta g(x;\boldsymbol{\xi})G(x;\boldsymbol{\xi})^{\alpha\theta-1}}{[1-G(x;\boldsymbol{\xi})^{\theta}]\{G(x;\boldsymbol{\xi})^{\alpha\theta}+[1-G(x;\boldsymbol{\xi})^{\theta}]^{\alpha}\}}. \tag{2.4}$$

One advantage of this approach is that if the baseline distribution has a closed-form expression (explicitly standard mathematical functions), the resulting representation of the

linear combination of the GOLL-G can be more mathematically tractable, making it easier to compute moments, gf and other measures.

Table 1 summarizes the associated sub-families of Equation (2.2). Furthermore, the log-likehood test makes it simple and interesting to compare the generalized class of distributions with specific classes of models.

Table 1 – Sub-families associated to the GOLL-G family of distributions.

$\boldsymbol{\alpha}$	$\boldsymbol{\theta}$	**Submodel**
-	1	Generalized log-logistic family (GLEATON; LYNCH, 2006)
1	-	Proportional reversed hazard rate family (GUPTA; GUPTA, 2007)
1	1	Baseline

2.2 MAIN PROPERTIES

The simple form of the exponentiated generator allows a broad spectrum of shapes, making it a well-suited alternative for diverse data types and phenomena, see (GUPTA; KUNDU, 1999). The distribution class *exponentiated*-G (Exp-G) is significant since it permits the properties of the GOLL-G family to be written from the exponentiated baseline. Such features enable it to be straightforward to compute the moments, generating function, among other measures and are used in some parts of the Section.

In this scenario, the Exp-G family for a parent cdf $G(x;\boldsymbol{\xi})$ is defined for a power extra parameter, c. Namely $W \sim$ Exp^c-G, the pdf and the cdf are given by $H_c(x) = G(x;\boldsymbol{\xi})^c$ and $h_c(x) = c\, g(x;\boldsymbol{\xi})G(x;\boldsymbol{\xi})^{c-1}$, respectively. Formally, for more than fifty baselines, several authors studied the characteristics of the Exp-G class, e.g., for Exp-log-normal (SHIRKE; S., 2006), for Exp-gamma (NADARAJAH; GUPTA, 2007) and for Exp-Gumbel (NADARAJAH, 2006), among others.

2.2.1 Linear Representation

The GOLL-G family of distributions admits the following linear representation:

$$f(y;\alpha,\theta,\boldsymbol{\xi}) = \sum_{k=0}^{\infty} b_k\, h_{k+1}(y;\boldsymbol{\xi}), \tag{2.5}$$

where b_k is

$$b_k = \frac{\alpha\theta}{k+1}\sum_{i,j=0}^{\infty}\sum_{l=k}^{\infty}(-1)^{j+k+l}\binom{-2}{i}\binom{l}{k}\binom{-(i+1)\alpha}{j}\binom{(i+1)\alpha\theta + j\theta - 1}{l},$$

used in subsequent Chapters to derive the linear representation of the proposed models and $h_{k+1}(y;\boldsymbol{\xi})$ is the Exp-G distribution with power parameter $k+1$.

Equation (2.5) is the main result of the Section defined on Cordeiro et al. (2017). By utilizing the properties of the Exp-G family of distributions, it is possible to derive significant properties of the GOLL-G class in a straightforward manner.

2.2.2 Quantile function

The quantile function (qf) of Y, denoted by $y = Q(u) = F^{-1}(u)$, can be obtained inverting Equation (2.2), as follows:

$$y = Q(u) = Q_{\mathsf{G}(x;\boldsymbol{\xi})}\left(\varepsilon_{\alpha,\theta}(u)\right), \tag{2.6}$$

where $Q_{\mathsf{G}(x;\boldsymbol{\xi})}(\cdot)$ is the qf of the baseline distribution and

$$\varepsilon_{\alpha,\theta}(u) = \left[\frac{\left(\frac{u}{1-u}\right)^{1/\alpha}}{1+\left(\frac{u}{1-u}\right)^{1/\alpha}}\right]^{1/\theta},$$

is utilized in the following Chapters.

Equation (2.6) is a useful tool for simulating any GOLL-G distribution when U is drawn from a uniform distribution on the interval $(0,1)$. Furthermore, the qf offers descriptive statistics, data exploration and more reliable estimation.

To effectively illustrate how skewness and kurtosis behave as functions of parameters, Galton's skewness (GALTON, 1883) and Moors kurtosis (MOORS, 1988) can be provide for a range of values of α and θ. To address these metrics, arbitrary values for $\boldsymbol{\xi}$ are utilized in the following Sections. These measures, which are considered using the quantile function, are less affected by outliers and exist even for distributions without moments.

Galton's skewness formula is

$$\mathcal{G} = \frac{Q(3/4) + Q(1/4) - 2\cdot Q(1/2)}{Q(3/4) - Q(1/4)},$$

whereas the Moors kurtosis is given by

$$\mathcal{M} = \frac{Q(3/8) - Q(1/8) + Q(7/8) - Q(5/8)}{Q(6/8) - Q(2/8)}.$$

2.2.3 Moments

Moments are a powerful tool for investigating the key features and characteristics of a distribution, such as tendency, dispersion, skewness and kurtosis. The definition below is used along with all applications to provide further information about any baseline proposed.

The nth moment of the GOLL-G distribution is defined in Cordeiro et al. (2017) by:

$$\mu'_n = \mathbb{E}(Y^n) = \sum_{k=0}^{\infty} b_k \ \mathbb{E}(H_{k+1}^n), \tag{2.7}$$

where can be employed the expressions for moments of some exponentiated distributions (NADARAJAH; KOTZ, 2006b). These expressions enable the computation of the GOLL-G moments, simplifying further analysis and characterization of the distribution.

2.2.4 Generating Function

The moment generating function (mgf) of Y, denoted as $M(t) = \mathbb{E}[e^{tY}]$, can be obtained by the following expression in Cordeiro et al. (2017), as:

$$M(t) = \sum_{k=0}^{\infty} b_k M_{k+1}(t), \tag{2.8}$$

where $M_{k+1}(t)$ represents the mgf of $H_{k+1}(y; \boldsymbol{\xi})$. The computation of the mgf by the Exp-G distribution reveals the properties of the GOLL-G distributions in a straightforward manner and is employed in all the applications of the proposed models.

2.3 ESTIMATION

Let $y_1, \ldots, y_n$ represent the observed values from the GOLL-G distribution with parameter vector $\boldsymbol{\psi} = (\alpha, \theta, \boldsymbol{\xi})^\top$ with dimension $r \times 1$. As defined in Cordeiro et al. (2017), the total log-likelihood function for $\boldsymbol{\psi}$ can be expressed, as follows:

$$\begin{aligned} l_n(\boldsymbol{\psi}) = n \log(\alpha\theta) + \sum_{i=1}^{n} \log[g(x_i; \boldsymbol{\xi})] + (\alpha\theta - 1) \sum_{i=1}^{n} \log[G(x_i; \boldsymbol{\xi})] \\ + (\alpha - 1) \sum_{i=1}^{n} \log[1 - G(x_i; \boldsymbol{\xi})^\theta] - 2 \sum_{i=1}^{n} \log\{G(x_i; \boldsymbol{\xi})^{\alpha\theta} + [1 - G(x_i; \boldsymbol{\xi})^\theta]^\alpha\}. \end{aligned} \tag{2.9}$$

The components of the score vector can be expressed, as:

$$\begin{aligned} U_\alpha &= \frac{n}{\alpha} + \theta \sum_{i=1}^{\infty} \log[G(x_i;\boldsymbol{\xi})] + \sum_{i=1}^{\infty} \log[1 - G(x_i;\boldsymbol{\xi})^\theta] \\ &- 2\sum_{i=1}^{\infty} \frac{(x_i;\boldsymbol{\xi})^{\alpha\theta} \log[G(x_i;\boldsymbol{\xi})] + [1 - G(x_i;\boldsymbol{\xi})^\theta]^\alpha \log[1 - G(x_i;\boldsymbol{\xi})^\theta]}{G(x_i;\boldsymbol{\xi})^{\alpha\theta} + [1 - G(x_i;\boldsymbol{\xi})^\theta]^\alpha}, \end{aligned}$$

$$\begin{aligned} U_\theta &= \frac{n}{\theta} + \alpha \sum_{i=1}^{\infty} \log[G(x_i;\boldsymbol{\xi})] + (1-\alpha) \sum_{i=1}^{\infty} \frac{G(x_i;\boldsymbol{\xi})^\theta \log[G(x_i;\boldsymbol{\xi})]}{1 - G(x_i;\boldsymbol{\xi})^\theta} \\ &- 2\alpha \sum_{i=1}^{\infty} \frac{G(x_i;\boldsymbol{\xi})^{\alpha\theta} \log[G(x_i;\boldsymbol{\xi})] - G(x_i;\boldsymbol{\xi})^\theta \log[1 - G(x_i;\boldsymbol{\xi})^\theta]^{\alpha-1} \log[G(x_i;\boldsymbol{\xi})]}{G(x_i;\boldsymbol{\xi})^{\alpha\theta} + [1 - G(x_i;\boldsymbol{\xi})^\theta]^\alpha}, \end{aligned}$$

and

$$\begin{aligned} U_{\boldsymbol{\xi}} &= \sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\xi}} g(x_i;\boldsymbol{\xi})}{g(x_i;\boldsymbol{\xi})} + (\alpha\theta - 1) \sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\xi}} G(x_i;\boldsymbol{\xi})}{G(x_i;\boldsymbol{\xi})} + \theta(1-\alpha) \sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\xi}} G(x_i;\boldsymbol{\xi}) G(x_i;\boldsymbol{\xi})^{\theta-1}}{1 - G(x_i;\boldsymbol{\xi})^\theta} \\ &- 2\sum_{i=1}^{\infty} \partial_{\boldsymbol{\xi}} G(x_i;\boldsymbol{\xi}) \frac{G(x_i;\boldsymbol{\xi})^{\alpha\theta-1} - G(x_i;\boldsymbol{\xi})^{\theta-1}[1 - G(x_i;\boldsymbol{\xi})^\theta]^{\alpha-1}}{G(x_i;\boldsymbol{\xi})^{\alpha\theta} + [1 - G(x_i;\boldsymbol{\xi})^\theta]^\alpha}, \end{aligned}$$

where $\partial_{\boldsymbol{\xi}} g(\cdot)$ and $\partial_{\boldsymbol{\xi}} G(\cdot)$ denotes the derivative of the function $g(x;\boldsymbol{\xi})$ and $G(x;\boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$, respectively.

The maximum likelihood estimate (MLE) $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ can be found by setting the score equations $U_\alpha = U_\theta = U_{\boldsymbol{\xi}} = 0$ using an iterative method algorithm to find roots. Alternatively, Equation (2.9) can be maximized numerically using the *optim* routine available in R Core Team (2021), or the *AdequacyModel* package (MARINHO et al., 2019) can serve as an alternative tool for this purpose.

In all applications throughout the Sections, the innovative model outperforms competitive, well-known generators (beta-G, Kumarawswamy-G, gamma-G, Weibull-G, among others) in terms of accuracy, insight and capture different non-normal patterns.

2.4 THE PROPOSAL GOLL-G REGRESSION MODEL

In the literature, location-scale and reparameterization regression models are presented in different forms, depending on the field in which they are applied. For the first one, Prataviera et al. (2019) defined the odd log-logistic geometric normal regression model to analyze medical data and Altun et al. (2023) introduced a new odd log-logistic family of distributions and applied the Weibull baseline to study medical heart data. In the second one, Prataviera et

al. (2019) used a reparameterization of the odd log-logistic generalized gamma distribution to introduce a new regression to analyze censored data and Cruz, Ortega e Cordeiro (2016) proposed the log-odd log-logistic regression model as a feasible alternative to investigating the well-known data of Stanford heart transplant.

In this context, a parametric regression model using the GOLL-G class of distributions, named the *GOLL-G regression model*, is offered as a viable alternative to the usual regression models, which requires no transformation or reparameterization. The most important part of parametric regression models is defining the parameters based on $\boldsymbol{x}$. In general, it is realized by considering the conditional expectation of the response variable in the presence of the covariates, which is useful for predicting the expectation value of Y. Nevertheless, in the case of this work, a link function is proposed to connect the vector parameter of the baseline distribution to the covariates.

The systematic component of the GOLL-G regression model takes into account the fact that the vector parameter $\boldsymbol{\xi}$ in Equation (2.3) varies between data ($i = 1, \ldots, n$) as described in:

$$\boldsymbol{\xi_i} = g(\boldsymbol{x}_i^\top \boldsymbol{\beta_j}), \tag{2.10}$$

where $g(\cdot)$ is a twice continuously differentiable link function (for example, identity, log, inverse, sqrt, etc) and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^\top$, for ($j = 1, \ldots, l$), is the parameter vector associated with the explanatory variables $\boldsymbol{x}_i^\top = (x_{i1}, \cdots, x_{ip})$. The components of $\boldsymbol{\beta_j}$ are assumed to be independent. Therefore, the function $g(\cdot)$ plays the link with the covariates and the new regression model.

Despite the simplicity of including explanatory variables, the model has some interpretability disadvantages. Due to the fact that baseline parameters are not measures of central tendency or variability, the results must be interpreted differently and it is proposed to associate them with skewness and/or kurtosis to analyze. However, the approach leaves open the possibility for future researchers to explore novel methods of linking variables that will assist in the understanding of the investigation and accurate its interpretability.

2.4.1 Estimation

In this Section, the form of the log-likelihood function of the proposed regression model is written in the similar pattern of the Equation (2.9), nonetheless each $\boldsymbol{\xi}_i$ will have a subindex

since it will be written using the Equation (2.10).

The components of the score vector of U_α and U_θ are the same presented in Cordeiro et al. (2017) and the component of the vector parameter $\boldsymbol{\xi_i}$ are defined to add the regression part, as follows:

$$\begin{aligned} U_{\boldsymbol{\xi_i}} &= \sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_j}} g(x_i;\boldsymbol{\xi_i})}{g(x_i;\boldsymbol{\xi_i})} + (\alpha\theta - 1)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_j}} G(x_i;\boldsymbol{\xi_i})}{G(x_i;\boldsymbol{\xi_i})} + \theta(1-\alpha)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_j}} G(x_i;\boldsymbol{\xi_i}) G(x_i;\boldsymbol{\xi_i})^{\theta-1}}{1 - G(x_i;\boldsymbol{\xi_i})^{\theta}} \\ &- 2\sum_{i=1}^{\infty} \partial_{\boldsymbol{\beta_j}} G(x_i;\boldsymbol{\xi_i}) \frac{G(x_i;\boldsymbol{\xi_i})^{\alpha\theta-1} - G(x_i;\boldsymbol{\xi_i})^{\theta-1}[1 - G(x_i;\boldsymbol{\xi_i})^{\theta}]^{\alpha-1}}{G(x_i;\boldsymbol{\xi_i})^{\alpha\theta} + [1 - G(x_i;\boldsymbol{\xi_i})^{\theta}]^{\alpha}}, \end{aligned}$$

where $\partial_{\boldsymbol{\beta_j}} g(\cdot) = \partial_{\boldsymbol{\xi_i}} g(\cdot)\partial_{\boldsymbol{\beta_j}}\boldsymbol{\xi_i}(\cdot)$ and $\partial_{\boldsymbol{\beta_j}} G(\cdot) = \partial_{\boldsymbol{\xi_i}} G(\cdot)\partial_{\boldsymbol{\beta_j}}\boldsymbol{\xi_i}(\cdot)$ denotes the derivatives of the parameter $\boldsymbol{\xi_i}$ using the chain rule.

The MLE $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ of the regression model is calculated setting the score equations $U_\alpha = U_\theta = U_{\boldsymbol{\xi_i}} = 0$ using an iterative method algorithm to find roots or using the *optim* routine.

Occasionally, it is difficult to obtain the MLE in a closed or semi-closed form, which arises from various factors: complexity of the likelihood function, lack of closed-form solutions, numerical issues, etc. Previous estimates from the baseline and data are used for each proposed model to figure out the initial guess of the parameters.

2.4.2 Accuracy of Maximum Likelihood Estimates

To assess the accuracy of the MLEs by means of Monte Carlo simulations, for some sample sizes, the average estimate (AE), bias, mean square error (MSE), absolute bias (AB), estimated average length (AL), root mean square error (RMSE) and/or coverage probability (CP) are computed for arbitrary parameter values (used in the following Sections). The measures are, as follows:

$$AE_\epsilon(n) = \frac{1}{N}\sum_{i=1}^{N}\hat{\epsilon}_i, \quad Bias_\epsilon(n) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\epsilon}_i - \epsilon), \quad MSE_\epsilon(n) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\epsilon}_i - \epsilon)^2,$$

$$AB_\epsilon(n) = \frac{1}{N}\sum_{i=1}^{N}|\hat{\epsilon}_i - \epsilon|, \quad AL_\epsilon(n) = \frac{3.919928}{N}\sum_{i=1}^{N} s_{\hat{\epsilon}_i}, \quad RMSE_\epsilon(n) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\epsilon}_i - \epsilon)^2}$$

and

$$CP_\epsilon(n) = \frac{1}{N}\sum_{i=1}^{N} I(\hat{\epsilon}_i - 1.95996 s_{\hat{\epsilon}_i}, \hat{\epsilon}_i + 1.95996 s_{\hat{\epsilon}_i}),$$

for $\boldsymbol{\epsilon} = (\alpha, \theta, \boldsymbol{\xi})$ and $s_{\hat{\epsilon}}$ are the standard errors (SEs) of the MLEs.

2.5 DIAGNOSTIC AND RESIDUAL ANALYSIS

To compare fitted models, some goodness-of-fit (GoF) statistics are commonly used. These include the Akaike information criterion (AIC), the consistent Akaike information criterion (CAIC), the Bayesian information criterion (BIC), given the complexity of the model (number of parameters), Cramér-von Mises (W^*), which compares an empirical distribution with a theoretical distribution, Anderson-Darling (A^*), which captures discrepancies in the tails of the distribution and Kolmogorov-Sminorv (KS), which compares the empirical distribution function with a reference probability distribution function. Generally, smaller values of these statistics indicate a better fit with the data.

The likelihood ratio (LR) statistic, which compares the likelihoods of two competing statistical models to see whether one provides a better explanation of the observed data, is used to evaluate the new distribution with its nested parameters to confirm the inclusion of the extra parameters (α and θ). This comparison helps determine whether the more complex model considerably improves the fit over the simpler one.

When examining outliers, a number of methods have been documented in the literature, see (COX; SNELL, 1968), (COOK; WEISBERG, 1982) and (ORTEGA; PAULA; BOLFARINE, 2008). In order to detect influential observations of a proposed regression model, diagnostic measures that involve excluding data points are employed. The objective of conducting residual analysis is to identify patterns or characteristics within the residuals, which could potentially affect the model's validity.

In this context, for any systematic component, the exclusion of observations follows:

$$\boldsymbol{\xi}_l = g(\boldsymbol{x_l}^\top \boldsymbol{\beta_j}), \quad l = 1, \ldots, n, \quad l \neq i. \tag{2.11}$$

For investigating the influential observations, the generalized Cook's distance is given by:

$$GCD_i = \left(\hat{\boldsymbol{\psi}}_{(i)} - \hat{\boldsymbol{\psi}}\right)^\top [\ddot{\boldsymbol{L}}(\hat{\boldsymbol{\psi}})] \left(\hat{\boldsymbol{\psi}}_{(i)} - \hat{\boldsymbol{\psi}}\right), \tag{2.12}$$

and the likelihood distance, as:

$$LD_i = 2\left\{l(\hat{\boldsymbol{\psi}}) - l(\hat{\boldsymbol{\psi}}_{(i)})\right\}, \tag{2.13}$$

where the subscript i denotes the observation deleted from the dataset and $\ddot{\boldsymbol{L}}(\boldsymbol{\psi})$ is the observed information matrix.

The deviance residuals have been commonly used to assess the goodness-of-fit of regression models (SILVA; ORTEGA; PAULA, 2011). Additionally, these measures can be applied to GOLL-G regression models to assess assumptions and detect the presence of outliers.

It follows that the deviance residuals for the GOLL-G regression are given by:

$$r_{D_i} = \text{sgn}(\hat{r}_{M_i}) \left\{-2[\hat{r}_{M_i} + \delta_i \log(\delta_i - \hat{r}_{M_i})]\right\}^{1/2}, \tag{2.14}$$

where

$$\hat{r}_{M_i} = \delta_i + \log\left[1 - F(y_i; \hat{\alpha}, \hat{\theta}, \hat{\boldsymbol{\xi}})\right], \tag{2.15}$$

are the martingale residuals, δ_i is the censoring indicator and sign$(\cdot)$ is the signal function with a value $+1$ if the argument is positive and -1 if the argument is negative.

Atkinson (1987) proposed the construction of envelopes to support the analysis of the residuals with normal probability plots. Confidence bands are simulated for these envelopes, and if the model gives a good fit, most of the points will lie randomly inside. The construction of these confidence bands can be calculated using the following steps:

(i) Calculate the r_{D_i}'s for the considered model;

(ii) Using the fitted model, the response variable is simulated (k samples);

(iii) Calculate the deviance residuals for each fitted model to the sample (for $j = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, n$);

(iv) Sort the n residuals to have $r_{D_{(i)j}}$'s for each group;

(v) Calculate descriptive statistics (mean, minimum and maximum) of the orderly residuals for each i;

(vi) Plot the residuals r_{D_i}'s versus the expected percentile of the standard normal and the descriptive statistics.

2.6 CONCLUDING REMARKS

The *generalized odd log-logistic-G* family of distributions, developed by Cordeiro et al. (2017), is reviewed in this Chapter. The family extends some common classes of distributions studied recently, such as the odd log-logistic (GLEATON; LYNCH, 2006) and proportional

reversed hazard rate (GUPTA; KUNDU, 2001) families. The main advantage of the GOLL-G over other competing generators is its ability to fit bimodal, asymmetric and heavy tail data with greater flexibility, which is applied in many research areas (biology, engineering, economy, epidemiology, extreme value theory, metereology, among others).

Some mathematical properties of the GOLL-G class of distribution are presented, such as a linear representation, the quantile function, the moments and the moment generating function. The maximum likelihood estimate and inference for model parameters are discussed. The GOLL-G regression model is introduced, which a link function is proposed to connect the vector parameter of the baseline distribution to the exploratory variables.

The MLEs are assessed using maximum likelihood and the parameters of the GOLL-G regression model were evaluated using Monte Carlo simulation, which demonstrated the estimators' consistency using some measures. Goodness-of-fit measures, diagnostic analysis, deviance residuals and envelope plots are used to support the new model's adequacy.

3 A NEW LINDLEY EXTENSION TO ANALYZE COVID-19 VACCINATION

RESUMO

A distribuição Lindley log-logística *odd* generalizada é introduzida e é proposto o modelo de regressão com uma componente sistemática de forma para investigar os elementos que explicam as taxas de imunização primária completa de COVID-19 de 254 condados no estado do Texas, Estados Unidos América (EUA). Algumas propriedades foram apresentadas, os parâmetros estimados por máxima verossimilhança e a acurácia investigada através de simulações. Para avaliar o ajuste do modelo, foram utilizados análise de diagnóstico e resíduos *deviance*. Ao nível do condado, o modelo proposto identificou fatores críticos que influenciam o parâmetro de forma, tais como o número de hospitais que comunicam a vacinação, a taxa de pobreza, o condado ser ou não uma área metropolitana, a percentagem da população com diploma do ensino médio, o acesso à internet e o número de indivíduos com doença cardíaca crónica. Os resultados indicam um bom ajuste, com uso potencial em esforços de imunização e como um modelo alternativo para trabalhos futuros em vários conjuntos de dados.

Palavras-chaves: COVID-19. Distribuição Lindley. Família log-logística *odd* generalizada. Máxima verossimilhança. Regressão. Simulação. Vacinação.

ABSTRACT

The generalized Lindley log-logistic distribution odd is introduced and the regression model with a systematic shape component is proposed to investigate the elements that explain the completed primary immunization rates of COVID-19 of 254 counties in the state of Texas, US. Some properties were presented, the parameters estimated by maximum likelihood and the accuracy investigated through simulations. Diagnostic and residual analysis were used to assess the model's fit. At the county level, the proposed model identified critical factors that influence the shape parameter, such as the number of hospitals reporting vaccination, the poverty rate, whether the county is a metropolitan area or not, the percentage of the population with a high school degree, internet access and the number of individuals with chronic heart disease. The results indicate a good fit, with potential use in immunization efforts and as an alternative model for future work on various data sets.

Keywords: COVID-19. Lindley distribution. Generalized odd log-logistic family. Maximum likelihood. Regression. Simulation. Vaccination.

3.1 INTRODUCTION

In recent years, the COVID-19 pandemic, caused by the new coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has had an enormous impact on practically every element of human life, producing major disruptions to healthcare systems, the economy and social structures all around the world. The advancements in the fight against the pandemic, particularly vaccination, were a critical tool for protecting individuals and communities from the virus and aiding in the prevention of its spread.

Vaccination efforts are ongoing around the world to counteract the COVID-19 epidemic. Over 13.5 billion[1] doses of COVID-19 vaccines were administered globally until November 2023. The federal government of the US has taken major steps to assure vaccination availability and accessibility, including subsidizing vaccine manufacture, distribution and administration. Vaccination rates have been higher among older individuals and healthcare workers, but efforts are ongoing to ensure that the vaccine is available to all eligible individuals. Vaccination actions are crucial to preventing the spread of the coronavirus and maintaining public health, despite hurdles such as vaccine reluctance and supply chain concerns.

According to Our World in Data[2], the US administered over 676 million doses of COVID-19 vaccines until November 2023, with more than 81% of the eligible population receiving at least one dose and more than 69% fully vaccinated. In terms of vaccination rates, the US is ahead of many other countries, but disparities in immunization coverage persist among age groups and neighborhoods. Vaccination rates vary between countries, with some still failing to acquire and distribute sufficient vaccines.

In this instance, the use of statistical techniques to analyze pandemic data has been widespread in the US and other countries. A comprehensive study by Hughes et al. (2021) examines the correlation between vaccination rates and social vulnerability at the county level, revealing significant disparities in vaccination coverage across counties. Despite limited data on vaccination safety and efficacy during pregnancy, a recent study by Razzaghi et al. (2021) found that vaccination coverage increased across all racial and ethnic groups during the study

[1] <https://ourworldindata.org/covid-vaccinations>
[2] <https://covid.cdc.gov/covid-data-tracker/#vaccine-delivery-coverage>

period. Other studies, (KRISS et al., 2022; ALBRECHT, 2022; REIMER et al., 2022), revealed a correlation with some determinant factors and the COVID-19 vaccination rate.

As detailed, the Lindley distribution has advantages over other competing models. Researchers have contributed significantly to the topic by introducing and analyzing numerous generalizations of the Lindley distribution. The study of the Lomax-Lindley distribution in lifetime data (TARVIRDIZADE, 2021), the perspective of the Lindley distribution on the unit interval (KARAKAYA et al., 2022), the application of the Marshall-Olkin Lindley distribution in reliability data (HAMEED; SAIEED, 2022) and the application of the modified-Lindley distribution in three real data sets (COŞKUN et al., 2022) are some notable examples of these generalizations. A novel extension of the Lindley distribution is presented to estimate survival rates using US life tables (MOHAMED; ALI; YOUSOF, 2023). At last, using Bayesian and non-Bayesian predictions for four applications based on real-world data sets, Elgarhy et al. (2023) investigated the truncated inverse power Lindley distribution.

Consequently, the Chapter aims to determine the factors that explain the COVID-19 vaccination rate by constructing a new regression model based on the new proposed *generalized odd log-logistic Lindley* (GOLLL) distribution. The GOLL-G family has applicability across various fields, highlighting its superiority over well-known generators. For example, Vasconcelos et al. (2022) proposed a parametric and a partially linear regression model called GOLLBS using the average pricing data collected from producers and wholesalers. In comparison to other regression models in the literature, the new model's flexibility provides a better fit. Similarly, Vasconcelos et al. (2022) proposed additive and parametric partial linear regressions using the generalized odd log-logistic log-normal (GOLLLN) distribution with climatological and an aerobic stability test data. Finding covariables with nonlinear effects on the response variable proved to be a crucial application of the model. Lastly, Prataviera et al. (2023) defined the GOLLMaxMix model to analyze COVID-19 Chinese data. According to the fitted model, the age group has a substantial impact on the lifetime of COVID-19 patients.

Several authors have studied the links between numerous factors that influence vaccination rates, such as demography, socioeconomic variables and comorbidities. The development of new models that reflect the non-Gaussian data is critical for overcoming COVID-19 research gaps. Because of the additional parameters, the new distribution can model a wide range of data shapes. The GOLLL regression proposed aims to be an effective model for finding the elements that influence vaccination and might be considered as an alternative for future studies to aid in vaccination efforts.

Thus, the analysis of the COVID-19 completed primary vaccination series at the county level in the state of Texas is the main focus of this Chapter. The goal is to look into how explanatory factors affect the response variable, with a particular emphasis on the effects of vaccination in the US. The objective of this research is to significantly add to the knowledge on the subject and offer insightful information on the variables influencing the vaccination rate.

The structure of the Chapter is as follows. The key characteristics of the GOLLL distribution are defined in Section 3.2. A linear representation is presented and a few of its mathematical features. The maximum likelihood estimation technique is applied and the estimators' accuracy is tested through some simulations in Section 3.3. In Section 3.4, a new GOLLL regression model with a systematic structure for the shape parameter is introduced and the estimators' consistency is assessed. In Section 3.5, an application of the proposed model to COVID-19 vaccination rate data is considered and its performance is compared with other models. Diagnostic analysis and deviance residuals confirmed that the new model is the best fit to explain the current data. In addition, the study presented valuable findings that support vaccination efforts. Section 3.6 provides a summary of the study's principal findings.

3.2 THE GOLLL DISTRIBUTION

The Lindley distribution is defined by the cdf and pdf (for $x > 0$), as:

$$G(x;\lambda) = 1 - \frac{1+\lambda+\lambda x}{1+\lambda}\mathrm{e}^{-\lambda x} \tag{3.1}$$

and

$$g(x;\lambda) = \frac{\lambda^2}{(1+\lambda)}(1+x)\mathrm{e}^{-\lambda x}, \tag{3.2}$$

respectively, with shape parameter $\lambda > 0$.

The new distribution, GOLLL, is defined by inserting Equation (3.1) on (2.2) and the cdf of the distribution is, as follows:

$$F(y;\alpha,\theta,\lambda) = \frac{\left[1 - \frac{1+\lambda+\lambda x}{1+\lambda}\mathrm{e}^{-\lambda x}\right]^{\alpha\theta}}{\left[1 - \frac{1+\lambda+\lambda x}{1+\lambda}\mathrm{e}^{-\lambda x}\right]^{\alpha\theta} + \left[1 - \left(1 - \frac{1+\lambda+\lambda x}{1+\lambda}\mathrm{e}^{-\lambda x}\right)^{\theta}\right]^{\alpha}}. \tag{3.3}$$

Let $Y \sim$ GOLLL(α,θ,λ) be a rv having cdf (3.3). By differentiating it, the pdf of Y reduces to:

$$f(y;\alpha,\theta,\lambda) = \frac{\frac{\alpha\,\theta\,\lambda^2}{(1+\lambda)}(1+x)\mathrm{e}^{-\lambda x}\left[1-\frac{1+\lambda+\lambda x}{1+\lambda}e^{-\lambda x}\right]^{\alpha\theta-1}\left[1-\left(1-\frac{1+\lambda+\lambda x}{1+\lambda}\mathrm{e}^{-\lambda x}\right)^{\theta}\right]^{\alpha-1}}{\left\{\left[1-\frac{1+\lambda+\lambda x}{1+\lambda}\mathrm{e}^{-\lambda x}\right]^{\alpha\theta}+\left[1-\left(1-\frac{1+\lambda+\lambda x}{1+\lambda}\mathrm{e}^{-\lambda x}\right)^{\theta}\right]^{\alpha}\right\}^2}. \tag{3.4}$$

Its hrf is easily found substituting Equation (3.2) and (3.1) into Equation 2.4, as:

$$\tau(y;\alpha,\theta,\lambda) = \frac{\alpha\theta\left(\frac{\lambda^2}{(1+\lambda)}(1+x)\mathrm{e}^{-\lambda x}\right)\left(1-\frac{1+\lambda+\lambda x}{1+\lambda}\mathrm{e}^{-\lambda x}\right)^{\alpha\theta-1}}{\left[1-\left(1-\frac{1+\lambda+\lambda x}{1+\lambda}\mathrm{e}^{-\lambda x}\right)^{\theta}\right]\left\{\left(1-\frac{1+\lambda+\lambda x}{1+\lambda}\mathrm{e}^{-\lambda x}\right)^{\alpha\theta}+[1-\left(1-\frac{1+\lambda+\lambda x}{1+\lambda}\mathrm{e}^{-\lambda x}\right)^{\theta}]^{\alpha}\right\}}. \tag{3.5}$$

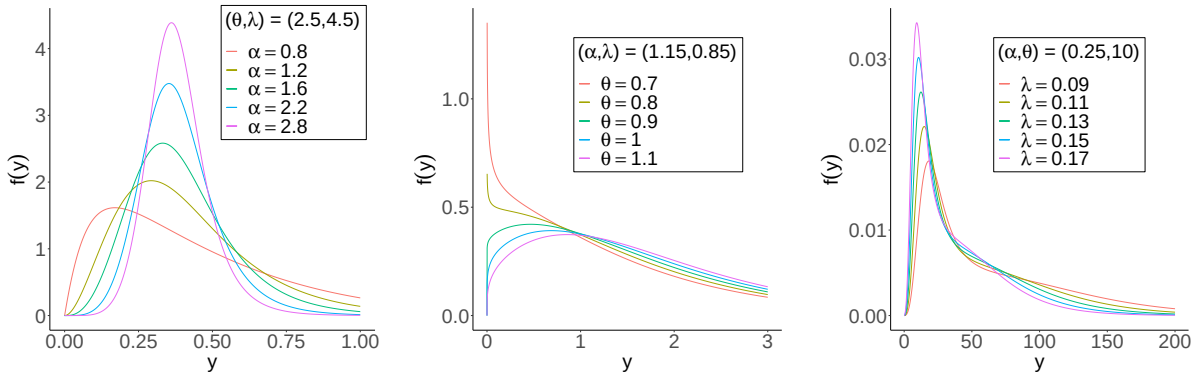
The special models of the GOLLL distribution, which demonstrate the broad applicability of the new class for fitting data throughout a range of distributions, are listed in Table 2.

Table 2 – Submodels associated to the GOLLL distribution.

α	θ	λ	**Submodel**
-	1	-	Odd log-logistic Lindley (OLLL) distribution (OZEL MORAD ALIZADEH; CANCHO, 2017)
1	-	-	Exponentiated-Lindley (Exp-L) distribution (NADARAJAH; BAKOUCH; TAHMASBI, 2011)
1	1	-	Lindley (L) distribution.

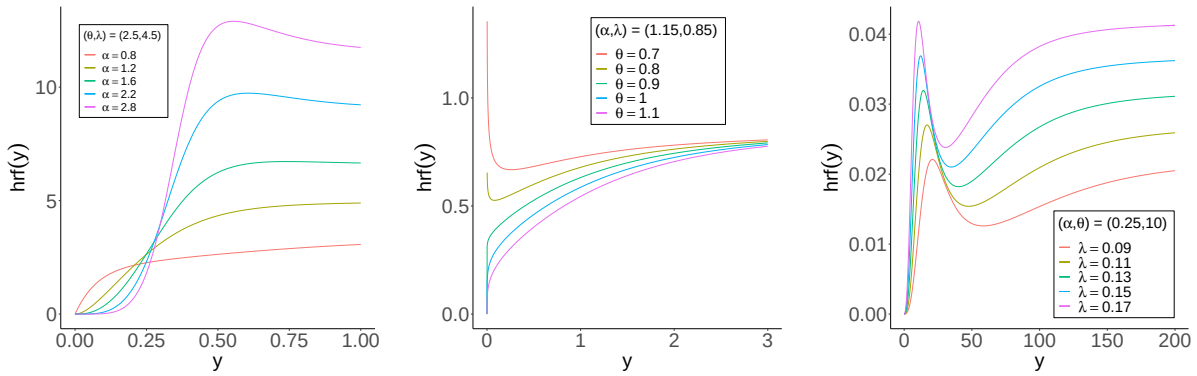
Plots of the pdf and the hrf of Y for chosen parameters are shown in Figures 1 and 2. The versatility of the GOLLL distribution in generating a wide range of hazard shapes is one of its most notable features, as presented by the extra shape parameters.

Figure 2 displays the new hrf model's inverse J-shape, increasing-decreasing, increasing-decreasing and other patterns and shapes. This is an improvement over the increasing Lindley hrf form. This versatility makes the model an effective tool for modeling varied data sets with different kinds of hazard rate patterns.



(a) (b) (c)

Figure 1 – GOLLL pdf for selected values.



(a) (b) (c)

Figure 2 – GOLLL hrf for selected values.

3.3 MAIN PROPERTIES

There are no mathematical closed-form properties of the GOLLL distribution. Therefore, utilizing the linear representation of Equation (2.5), the pdf of Y can be expressed, as:

$$f(y;\alpha,\theta,\lambda) = \sum_{k=0}^{\infty} b_k h_{k+1}(y;\lambda), \tag{3.6}$$

where $h_{k+1}(y;\lambda)$ is the Exp-L distribution with power parameter $k+1$ defined by:

$$h_{k+1}(y;\lambda) = (k+1)\left[\frac{\lambda^2}{(1+\lambda)}(1+x)\mathrm{e}^{-\lambda x}\right]\left[\frac{\lambda^2}{(1+\lambda)}(1+x)\mathrm{e}^{-\lambda x}\right]^k.$$

Thus, the GOLLL properties is obtained in a straightforward way by utilizing Exp-L properties, see (NADARAJAH; BAKOUCH; TAHMASBI, 2011).

3.3.1 Quantile function

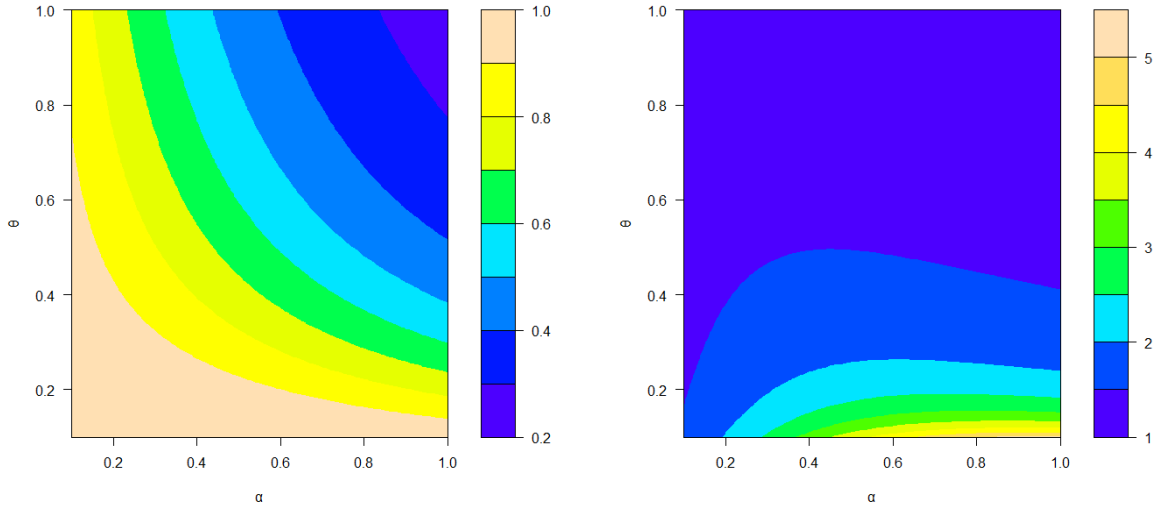
The qf of Y can be obtained from Equation (2.6), as:

$$Q(u) = -1 - \frac{1}{\lambda} - \frac{1}{\lambda}W_{-1}[(1+\lambda)(\varepsilon_{\alpha,\theta}(u)-1)\mathrm{e}^{-1-\lambda}], \tag{3.7}$$

where $W_{-1}(\cdot)$ denotes the negative branch of the Lambert W function[3].

Figure 3 displays Galton's skewness and Moors' kurtosis varying α and θ, with $\lambda = 5.25$. These plots highlight the impact of both parameters on the distribution shape. When both parameters α and θ increase, the skewness and kurtosis measures drop to a minimum region.

[3] <https://mathworld.wolfram.com/LambertW-Function.html>



(a) (b)

Figure 3 – GOLLL distribution. (a) Galton's skewness. (b) Moors' kurtosis.

3.3.2 Moments

The moments of the GOLLL distribution are obtainable considering the Exp-L distribution in Equation (2.7).

Theorem 1. *The nth ordinary moment of Y is given by:*

$$\mu'_n = \mathbb{E}(Y^n) = \sum_{k=0}^{\infty} \frac{(k+1)\lambda^2}{1+\lambda} K(k+1, \lambda, n, \lambda)\, b_k, \tag{3.8}$$

where

$$K(a,b,c,\delta) = \sum_{i=0}^{\infty}\sum_{j=0}^{i}\sum_{l=0}^{j+1} \binom{a-1}{i}\binom{i}{j}\binom{j+1}{l} \frac{(-1)^i\, b^j\, \Gamma(1+l+c)}{(1+b)^i\,(bi+\delta)^{1+c+l}}.$$

Proof. The proof is straightforward by applying Equation (2.5) and using the Exp-L moments, see (NADARAJAH; BAKOUCH; TAHMASBI, 2011). □

3.3.3 Generating function

The mgf of the GOLLL distribution is computed considering the gf of the Exp-L distribution in Equation (2.8).

Theorem 2. *The mgf of the GOLLL density can be expressed, as:*

$$M_Y(t) = \sum_{k=0}^{\infty} \frac{(k+1)\lambda^2}{1+\lambda} K(k+1, \lambda, 0, \lambda - t)\, b_k, \quad \textit{for} \quad t < \lambda.$$

Proof. The proof is straightforward applying the Theorem (2.5) and using the mgf of the Exp-L, see (NADARAJAH; BAKOUCH; TAHMASBI, 2011) and (RANJBAR; ALIZADEH; ALTUN, 2019). □

3.3.4 Estimation

Let $y_1, \ldots, y_n$ be a random sample from $Y \sim \text{GOLLL}(\alpha, \theta, \lambda)$. For the parameter vector $\boldsymbol{\psi} = (\alpha, \theta, \lambda)^\top$, the log-likelihood function is provided by:

$$\begin{aligned} l_n(\boldsymbol{\psi}) = {} & n\log(\alpha\theta) + \sum_{i=1}^{n}\log\left[\frac{\lambda^2}{(1+\lambda)}(1+y_i)\mathrm{e}^{-\lambda y_i}\right] + (\alpha\theta-1)\sum_{i=1}^{n}\log\left[1-\frac{1+\lambda+\lambda y_i}{1+\lambda}\mathrm{e}^{-\lambda y_i}\right] \\ & + (\alpha-1)\sum_{i=1}^{n}\log\left[1-\left(1-\frac{1+\lambda+\lambda y_i}{1+\lambda}\mathrm{e}^{-\lambda y_i}\right)^{\theta}\right] \\ & - 2\sum_{i=1}^{n}\log\left\{\left(1-\frac{1+\lambda+\lambda y_i}{1+\lambda}\mathrm{e}^{-\lambda y_i}\right)^{\alpha\theta} + \left[1-\left(1-\frac{1+\lambda+\lambda y_i}{1+\lambda}\mathrm{e}^{-\lambda_i y_i}\right)^{\theta}\right]^{\alpha}\right\}. \end{aligned} \tag{3.9}$$

For simplicity, let

$$A_i = A_i(\lambda) = \left(1-\frac{1+\lambda+\lambda y_i}{1+\lambda}\mathrm{e}^{-\lambda y_i}\right).$$

Then, the components of the score vector are

$$\begin{aligned} U_\alpha &= \frac{n}{\alpha} + \theta\sum_{i=1}^{n}\log(A_i) + \sum_{i=1}^{n}\log\left(1-{A_i}^{\theta}\right) \\ &- 2\sum_{i=1}^{n}\frac{\theta\log(A_i){A_i}^{\alpha\theta} + (1-{A_i}^{\theta})^{\alpha}\log(1-{A_i}^{\theta})}{{A_i}^{\alpha\theta} + (1-{A_i}^{\theta})^{\alpha}}, \end{aligned}$$

$$\begin{aligned} U_\theta &= \frac{n}{\theta} + \alpha\sum_{i=1}^{n}\log(A_i) - (\alpha-1)\sum_{i=1}^{n}\frac{{A_i}^{\theta}\log(A_i)}{1-{A_i}^{\theta}} \\ &+ \sum_{i=1}^{n}\frac{\alpha{A_i}^{\alpha\theta}\log(A_i) + (1-{A_i}^{\alpha})^{\theta}\log(1-{A_i}^{\alpha})}{{A_i}^{i\alpha\theta} + (1-{A_i}^{\alpha})^{\theta}} \end{aligned}$$

and

$$U_\lambda = n\left(\frac{\lambda+2}{\lambda^2+\lambda}\right) - \sum_{i=1}^{n}y_i - (\alpha\theta-1)\sum_{i=1}^{n}\frac{\lambda y_i(2+\lambda+y_i+\lambda y_i)}{(1+\lambda)[1+\lambda-e^{\lambda y_i}(1+\lambda)+\lambda y_i]}.$$

Setting the score equations $U_\alpha = U_\theta = U_\lambda = 0$, the MLE of $\boldsymbol{\psi}$ can be found using a Newton-Raphson type algorithm. Alternatively, the *optim* routine can be used to numerically maximize Equation (3.9).

3.3.5 Simulation study

To assess the accuracy of the estimators, $1,000$ samples of sizes $50, 150, 300, 500, 750,$ and $1,000$ are generated under two scenarios: $\boldsymbol{\psi} = (0.50, 0.58, 0.77)^\top$ for scenario 1, and $\boldsymbol{\psi} = (0.72, 0.80, 0.95)^\top$ for scenario 2. For every sample size, for $\boldsymbol{\epsilon} = (\alpha, \theta, \lambda)$, the AEs, biases and MSEs are reported in Table 3.

Table 3 – Simulations results for GOLLL distribution.

Scenario 1									
Par	**n = 50**			**n = 150**			**n = 300**		
	AE	**Bias**	**MSE**	**AE**	**Bias**	**MSE**	**AE**	**Bias**	**MSE**
α	0.525	0.025	0.157	0.527	0.027	0.064	0.504	0.004	0.030
θ	1.059	0.479	1.002	0.733	0.153	0.212	0.679	0.099	0.102
λ	1.203	0.433	0.919	0.913	0.143	0.258	0.861	0.091	0.125
Par	**n = 500**			**n = 750**			**n = 1000**		
	AE	**Bias**	**MSE**	**AE**	**Bias**	**MSE**	**AE**	**Bias**	**MSE**
α	0.499	-0.001	0.018	0.506	0.006	0.012	0.508	0.008	0.009
θ	0.644	0.064	0.051	0.611	0.031	0.029	0.597	0.017	0.020
λ	0.833	0.063	0.068	0.798	0.028	0.041	0.784	0.014	0.029
Scenario 2									
Par	**n = 50**			**n = 150**			**n = 300**		
	AE	**Bias**	**MSE**	**AE**	**Bias**	**MSE**	**AE**	**Bias**	**MSE**
α	0.936	0.186	0.794	0.859	0.109	0.289	0.803	0.053	0.105
θ	1.541	0.721	2.924	1,057	0.237	0.516	0.922	0.102	0.189
λ	1.296	0.346	0.968	1.073	0.123	0.313	1.000	0.050	0.142
Par	**n = 500**			**n = 750**			**n = 1000**		
	AE	**Bias**	**MSE**	**AE**	**Bias**	**MSE**	**AE**	**Bias**	**MSE**
α	0.794	0.044	0.067	0.772	0.022	0.039	0.772	0.011	0.013
θ	0.876	0.056	0.111	0.867	0.047	0.074	0.844	0.028	0.043
λ	0.973	0.023	0.092	0.972	0.022	0.062	0.958	0.028	0.087

The consistency criteria is held by the results, which show that the AEs converge to the true parameters and that the biases and MSEs decrease as sample size increases. The results for scenarios 1 and 2, for all parameter estimates obtained at $n = 50$, are superestimates, indicating sensitivity to specific values and sample sizes. Nevertheless, when n increases, the estimatives approach real values.

3.4 THE GOLLL REGRESSION MODEL

The systematic component of the GOLLL regression model takes into account the fact that the parameter λ in Equation (3.4) varies across observations ($i = 1, \ldots, n$), as:

$$\lambda_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \tag{3.10}$$

where is a twice continuously differentiable log-linear link function and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top$ is the p-dimensional parameter vector associated with the explanatory variables $\boldsymbol{x}_i^\top = (x_{i1}, \cdots, x_{ip})$.

3.4.1 Estimation

The components of the score vector of U_α and U_θ are the same Equations presented in Subsection (3.3.4) and the score component of the vector parameter λ_i is defined to add the regression part, as follows:

$$\begin{aligned} U_{\lambda_i} &= \sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta}} g(x_i;\lambda_i)}{g(x_i;\lambda_i)} + (\alpha\theta - 1)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta}} G(x_i;\lambda_i)}{G(x_i;\lambda_i)} + \theta(1-\alpha)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta}} G(x_i;\lambda_i) G(x_i;\lambda_i)^{\theta-1}}{1 - G(x_i;\lambda_i)^{\theta}} \\ &- 2\sum_{i=1}^{\infty} \partial_{\boldsymbol{\beta}} G(x_i;\lambda_i) \frac{G(x_i;\lambda_i)^{\alpha\theta-1} - G(x_i;\lambda_i)^{\theta-1}[1 - G(x_i;\lambda_i)^{\theta}]^{\alpha-1}}{G(x_i;\lambda_i)^{\alpha\theta} + [1 - G(x_i;\lambda_i)^{\theta}]^{\alpha}}, \end{aligned}$$

where $\partial_{\boldsymbol{\beta}} g(x_i;\lambda_i) = \partial_{\lambda_i} g(x_i;\lambda_i)\partial_{\boldsymbol{\beta}}\lambda_i(x_i;\boldsymbol{\beta})$ and $\partial_{\boldsymbol{\beta}} G(x_i;\lambda_i) = \partial_{\lambda_i} G(x_i;\lambda_i)\partial_{\boldsymbol{\beta}}\lambda_i(x_i;\boldsymbol{\beta})$ denotes the derivatives of the parameter λ_i using the chain rule.

The MLE $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ of the regression model is calculated setting the score equations $U_\alpha = U_\theta = U_{\lambda_i} = 0$ using an iterative method algorithm to find roots or using the *optim* routine.

3.4.2 Simulation study

The accuracy of the MLEs in the GOLLL regression model can be assessed using the measures: biases, MSEs, ALs and CPs. Equation (3.7) is used to generate one-thousand samples of sizes $n = 25, 55, \ldots, 1,000$ by setting $\alpha = 0.75$, $\theta = 1.85$, $\beta_0 = 2.75$ and $\beta_1 = 3.40$. The Monte Carlo simulation offers a suitable method for examining the model's parameters, allowing researchers to investigate how a distribution behaves in different scenarios.

Figures 4-10 report how the measures evaluate with the sample size. The biases, MSEs and ALs decay toward zero when sample size increases. It is also possible to detect, in the biases plots, that the parameters β_0 and β_1 were underestimated due to the model's complexity or

numerical issues, yet, as n rises, convergence is checked. Moreover, the CPs approach the value of 0.95 when n increases. These findings provide strong evidence of the consistency of the MLEs.

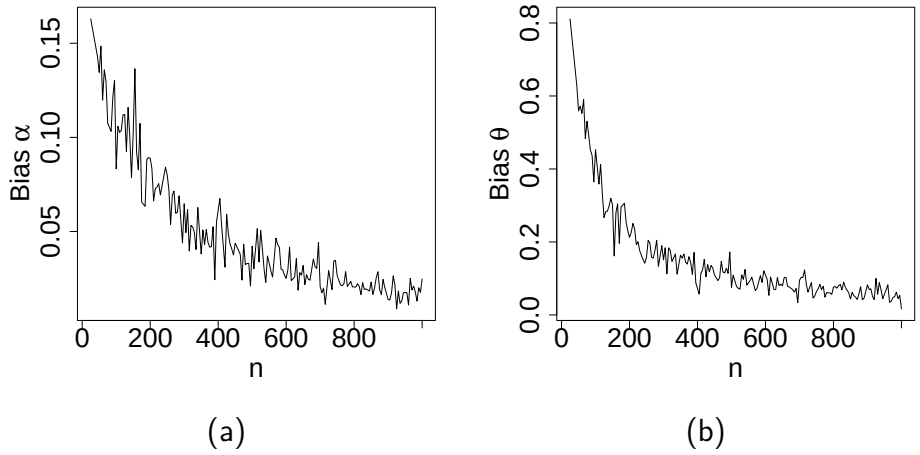


Figure 4 – Biases versus sample size from GOLLL regression model.

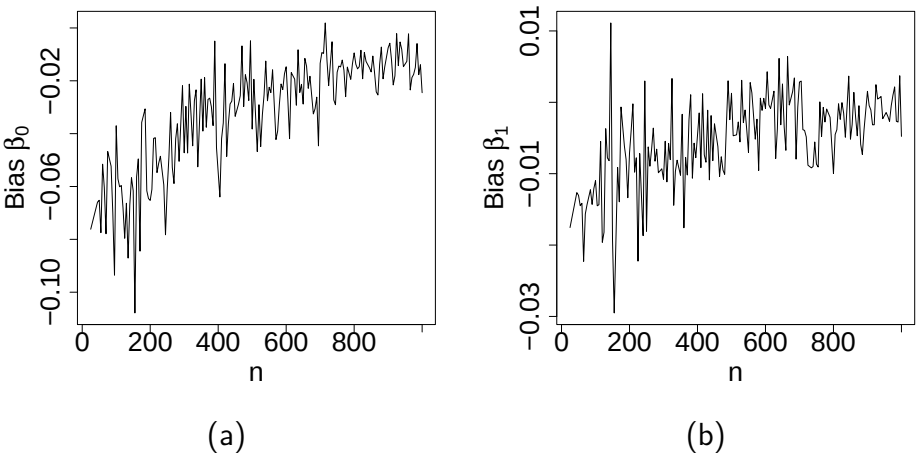
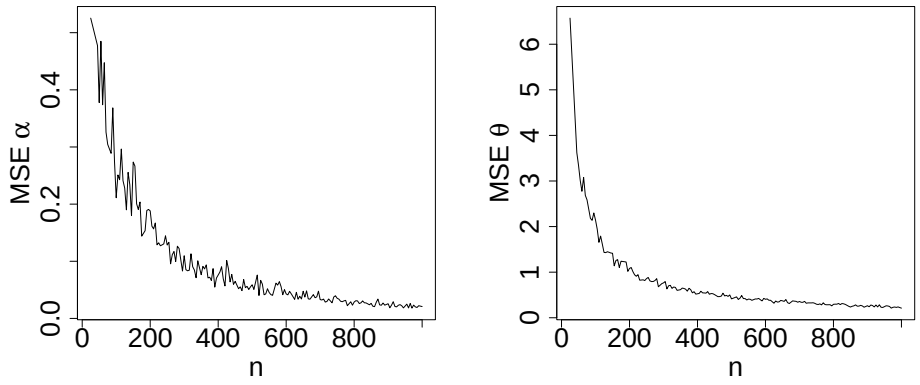
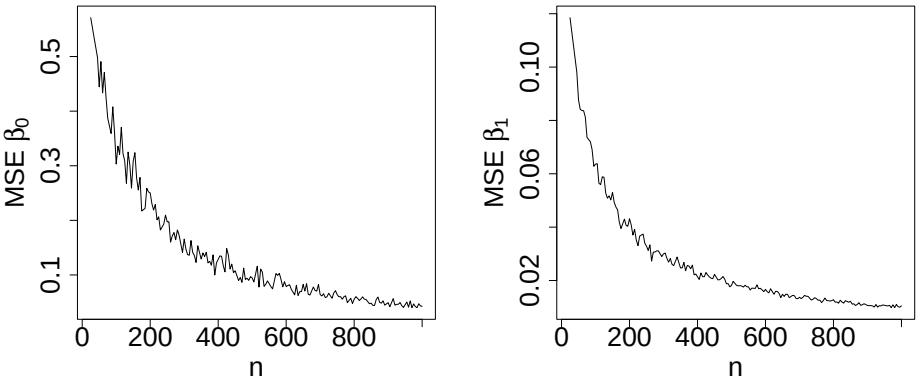


Figure 5 – Biases versus sample size from GOLLL regression model.



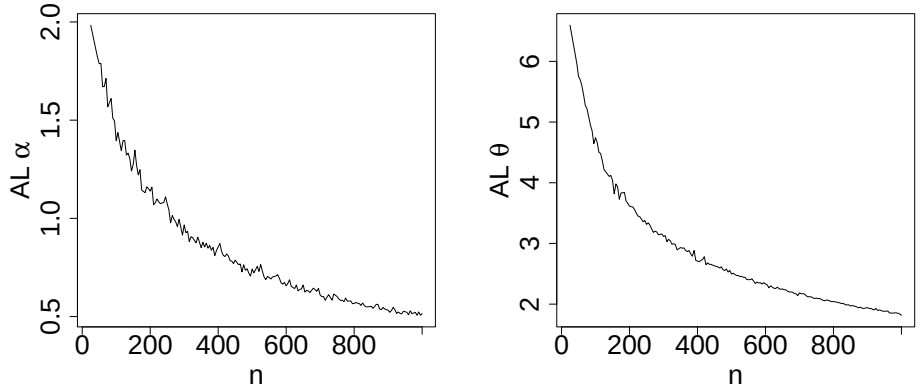
(a) (b)

Figure 6 – MSEs versus sample size from GOLLL regression model.



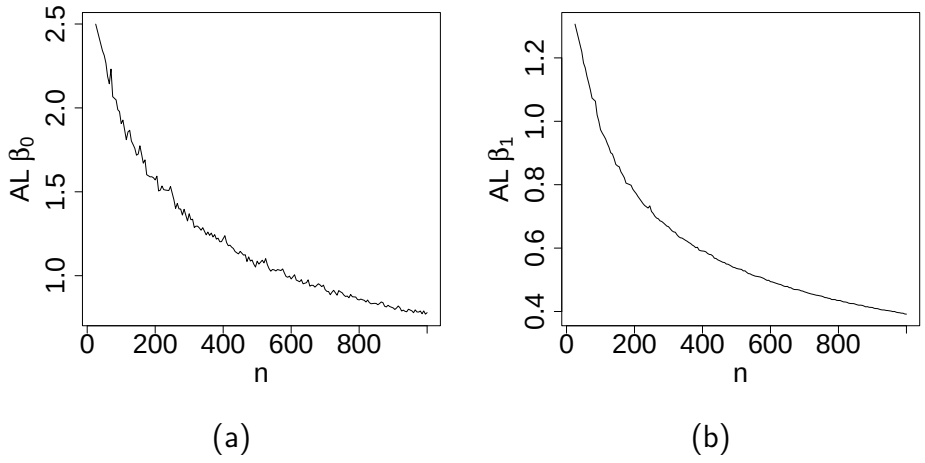
(a) (b)

Figure 7 – MSEs versus sample size from GOLLL regression model.



(a) (b)

Figure 8 – ALs versus sample size from GOLLL regression model.



(a) (b)

Figure 9 – ALs versus sample size from GOLLL regression model.

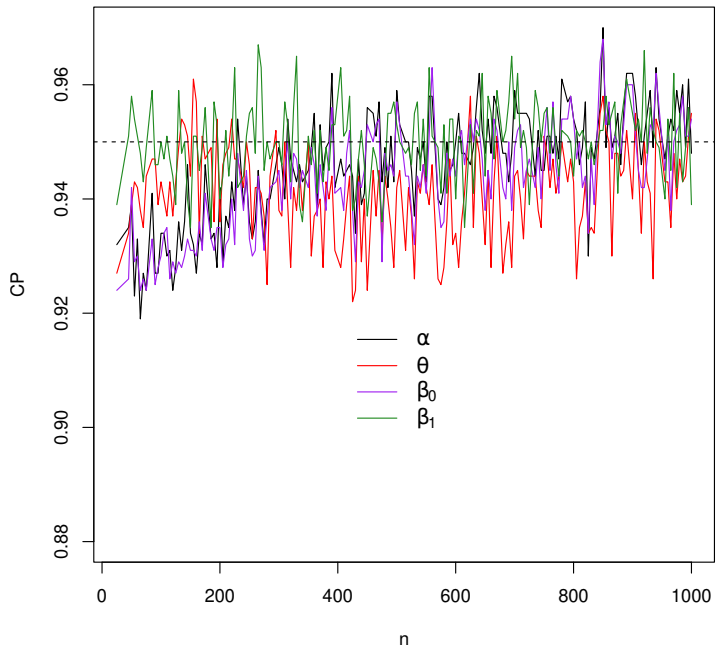


Figure 10 – CPs versus sample size from GOLLL regression model.

3.5 APPLICATION: TEXAS COUNTIES COVID-19 VACCINATION RATE DATA

Table 4 shows alternate distributions of well-known generators, which were compared to the GOLLL model and the nested models. The distributions are provided (for $x > 0$), respectively:

$$F_{\text{BL}}(x) = I_{G(x)}(a,b) = \frac{1}{B(a,b)} \int_0^{G(x)} w^{a-1}(1-w)^{b-1} dw,$$

Table 4 – Competitive distributions compared to the GOLLL distribution.

Distribution	Reference
Beta-Lindley (BL)	(MEROVCI; SHARMA et al., 2014)
Kumaraswamy -Lindley (KwL)	(ÇAKMAKYAPAN; KADILAR, 2014)
Gamma-Lindley (GL)	(ZEGHDOUDI; NEDJAR, 2016)
Weibull-Lindley (WL)	(ASGHARZADEH; NADARAJAH; SHARAFI, 2016)

$$F_{\mathsf{KwL}}(x) = 1 - [1 - G(x)^a]^b,$$
$$F_{\mathsf{GL}}(x) = \frac{\gamma\{a, -\log[1 - G(x)]/b\}}{\Gamma(a)}$$

and

$$F_{\mathsf{WL}}(x) = 1 - \exp[a\, G(x)],$$

where the parameters of all distributions are positive real numbers, $G(x)$ is defined in Equation (3.1) and $I_{G(x)}(a,b) = \frac{1}{B(a,b)} \int_0^{G(x)} w^{a-1}(1-w)^{b-1} dw$ is the incomplete beta function. For all fitted models, the *goodness.fit* function from *AdequacyModel*, using the BFGS method computes the MLEs (SEs in parentheses).

3.5.1 A data set definition

An application is presented using the completed primary series county-level COVID-19 immunization rates in the state of Texas, US, to illustrate the utility of the novel GOLLL regression model over competing models.

Texas is used because it has the largest number of counties of all states and has a large national impact on industries such as oil, coal, natural gas, nuclear energy and is the main generator of wind energy in US. Moreover, the state's public health system throughout the pandemic mitigating the spread of COVID-19 in its early stages using local authorities to enforce stay-at-home orders, adding to work of the state's health care personnel. In contrast, the pandemic showed the underfunding of local, state and federal health systems[4].

The proposed model has advantages and limitations over unitary models. The Lindley distribution used as a baseline has certain interesting characteristics, including flexibility in modeling asymmetric data, simplicity which allows it easy to apply, versatility for count data, and usefulness in survival models, among others. Among the issues that can arise while applying the model are problems with the results' interpretability, forecasts outside the range of the

[4] <https://www.texastribune.org/2021/06/11/watch-coronavirus-1623357664/>

response variable, modeling bias, and model sensitivity, among others. Despite this, the model captures the impact of the covariates on the response variable very effectively.

The data set refers to 254 percentages of the population in counties with a completed vaccination (aged adjust) to COVID-19 extracted from CDC[5] (accessed on 22nd February 2023). To assess the accuracy of the new model, more research utilizing different data sets (countries, states, and nations) ought to be conducted.

The explanatory variables were extracted from County Health Rankings[6] (data from 2020) (accessed on 22nd February 2023) are outlined (for $i = 1, \ldots, 254$) below:

1. y_i: Population rate with complete primary series of COVID-19 vaccination (VR) (response variable);
2. x_{i1}: Total number of hospitals reporting vaccination (HP);
3. x_{i2}: Poverty rate (PR) (percentage of individuals with income below the poverty line);
4. x_{i3}: Metropolitan status (MS) ($0 =$ metropolitan, $1 =$ non-metropolitan);
5. x_{i4}: High school completion rate (HR) (proportion of individuals aged 25 and above who have completed high school or its equivalent);
6. x_{i5}: Broadband access (BA) (percentage of households that have access to broadband internet);
7. x_{i6}: Heart disease rate (HT) (percentage of individuals that have chronic heart disease).

Table 5 provides descriptive statistics for the data set. The vaccination rate follows a right-skewed distribution with heavy tails, showed in Figure 11. Some hypotheses can be formulated, as follows:

- The lower VR mean (0.483), compared to the total population completed primary vaccinated in the US (0.695[7]), is explained by some factors, such as the social vulnerability of some counties in the State (MOFLEH et al., 2022), demographics and psychosocial variables (ethnicity, education, gender identification, age, trust, fear, etc) (LUNINGHAM et al., 2023) and mainly vaccine hesitancy among the population and healthcare personality (BERRY; ADAMS; VYTLA, 2024) and (HOSEK et al., 2022);

[5] <https://covid.cdc.gov/covid-data-tracker/#datatracker-home>
[6] <https://www.countyhealthrankings.org/>
[7] <https://covid.cdc.gov/covid-data-tracker/#vaccinations_vacc-people-booster-percent-pop5>

Table 5 – Descriptive statistics of COVID-19 completed primary vaccination rates data.

Variable	Statistics						
	Mean	**Median**	**SD**	**Skewness**	**Kurtosis**	**Min.**	**Max.**
VR	0.483	0.452	0.132	1.485	6.021	0.189	0.950
HP	1.717	1.000	4.282	6.666	56.447	0.000	45.000
PR	0.161	0.152	0.061	1.022	4.878	0.026	0.395
HR	0.818	0.830	0.085	-2.056	12.509	0.220	0.970
BA	0.769	0.770	0.084	-0.388	3.301	0.480	0.970
HT	0.082	0.082	0.017	0.248	2.906	0.045	0.134

- The HP could result in increased VR. Further investigations is required;
- Lower vaccination rates are expected in high-poverty communities. COVID-19 increased poverty and inequality worldwide (BUHEJI et al., 2020; DEATON, 2021). Individuals living in poverty may lack access to reliable transportation, face barriers to accessing health-care facilities and have limited resources for paying out-of-pocket costs associated with vaccination (HYDER et al., 2021; PAROLIN; LEE, 2022). The study of Liao (2021) revealed the lack of access to the COVID-19 vaccine in the lowest county's poverty rates across the American state of Illinois. Another study, (OLIVEIRA et al., 2021), showed a strong negative correlation with poverty and vaccine coverage in the 189 countries' research. This can result in lower vaccination rates among populations living in poverty;
- Lower VR are expected in non-metropolitan areas. The differences in vaccination rates between urban and rural communities are likely driven by various factors, such as differences in access to healthcare resources, vaccine distribution challenges and mainly vaccine hesitancy (MURTHY et al., 2021). Patterns in COVID-19 vaccination coverage by urbanity are addressed by Barry et al. (2021), which indicated lower vaccination rates in rural than urban area. Further, disparities in COVID-19 vaccination coverage between urban and rural counties and explained them by educational attainment, healthcare infrastructure and Trump vote share (SUN; MONNAT, 2022);
- Furthermore, higher HR may indicate a better comprehension of the need of immunization. Thus, counties with higher high school graduation rates tend to have higher vaccination rates as well, which can be attributed to more access to accurate information regarding vaccines to access better healthcare and vaccination services (KHAIRAT; ZOU; ADLER-MILSTEIN, 2022). Other studies, (MALIK et al., 2020; AGARWAL et al., 2021;

COUGHENOUR et al., 2021), revealed that high school is a key difference in coverage, access and hesitancy vaccination;

- Locations with a higher BA indicate greater access to vaccination-related information. Websites and social media platforms have been used to disseminate information about vaccine availability, eligibility and safety. The study's results suggest that counties with greater access to broadcast media have a higher COVID-19 vaccination rate, which highlights the disparities in access to the internet and technology among some communities. This finding is consistent with other research (GOEL; NELSON, 2021). Alternative studies, showed that lack of internet access is a barrier to vaccination. In New York City and some counties in North Carolina, the COVID-19 vaccine hesitancy increases if there is difficulty accessing the internet (MICHAELS; PIRANI; CARRASCAL, 2021; DOHERTY et al., 2021);

- Finally, counties with higher HT may prioritize immunization. Several studies (CLERKIN et al., 2020; TIPIRNENI et al., 2022; GUAN et al., 2020) have demonstrated the heightened risk of individuals with chronic heart disease contracting and experiencing severe symptoms from COVID-19, as well as increased rates of hospitalization and mortality. For these reasons, many states in the US have implemented targeted outreach efforts to ensure that these populations have access to the vaccine. Hence, the study's results indicate that counties with high rates of chronic heart disease have a correspondingly higher rate of vaccination. This finding highlights the importance of the government's focus on prioritizing at-risk populations (OSUAGWU et al., 2022). Subsequent studies, (CHOI; CHEONG, 2021; YELIN et al., 2021), illustrated the efficacy and safety of the COVID-19 vaccine based on the presence of comorbidities, including heart disease.

3.5.2 Results

First, the analysis solely models the response variable using the nested models and BL, KwL, GL and WL distributions. Table 6 reports the MLEs, SEs and GoF statistics (with the p-values of KS) for the fitted distributions to the COVID-19 vaccination rate data. Based on these measures, the GOLLL distribution is the best fit for the available data.

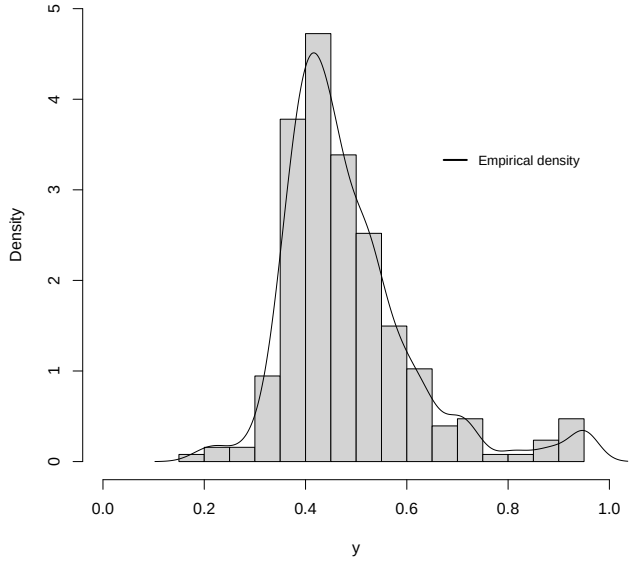


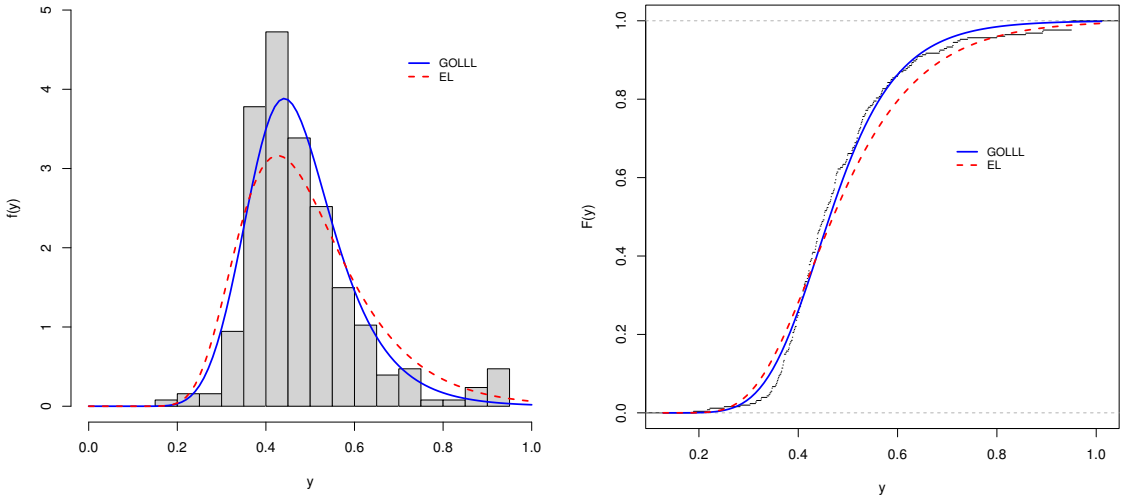
Figure 11 – Histogram and empirical density of COVID-19 completed primary vaccination rates data.

Table 6 – Findings from the fitted models of COVID-19 completed primary vaccination rates data.

Model	Parameters			W^*	A^*	KS
GOLLL$(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\lambda})$	**1.490**	**18.003**	**7.814**	**0.306**	**2.076**	**0.059**
	(0.0002)	**(0.0001)**	**(0.0003)**			**(0.338)**
OLLL(α, λ)	4.985	1	2.084	0.444	2.774	0.067
	(0.264)	(-)	(0.025)			(0.202)
EL(θ, λ)	1	35.255	9.127	0.321	2.275	0.097
	(-)	(5.861)	(0.400)			(0.015)
L(λ)	1	1	2.640	0.702	4.480	0.450
	(-)	(-)	(0.136)			($<$0.0001)
BL(a, b, λ)	30.015	1.810	7.245	0.397	2.696	0.079
	(7.324)	(0.503)	(1.155)			(0.082)
KwL(a, b, λ)	20.113	2.156	6.676	0.493	3.235	0.081
	(4.372)	(0.514)	(0.764)			(0.069)
GL(a, b, λ)	7.807	0.005	0.307	0.793	5.013	0.126
	(0.751)	($<$0.001)	(0.022)			($<$0.0001)
WL(a, λ)	121.307	0.124		0.815	5.140	0.436
	(44.059)	(0.024)				($<$0.0001)

Figure 12(a) illustrates the histogram and fitted densities of the two best models. The estimated cdfs of these models are also shown in Figure 12(b) The Figures corroborate the findings and the proposed distribution fits the current data well.

Three LR tests compare the GOLLL distribution with its nested models. The findings in Table 7 indicate that the inclusion of extra parameters is significant for accurately modeling the current data.



(a) (b)

Figure 12 – Fitted models of COVID-19 completed primary vaccination rates data. (a) Histogram and estimated pdfs. (b) Empirical and estimated cdfs.

Table 7 – LR tests of the GOLLL distribution.

Models	**Statistic w**	**p-value**
GOLLL vs L	498.805	< 0.0001
GOLLL vs EL	15.127	0.0001
GOLLL vs OLLL	5.194	0.0227

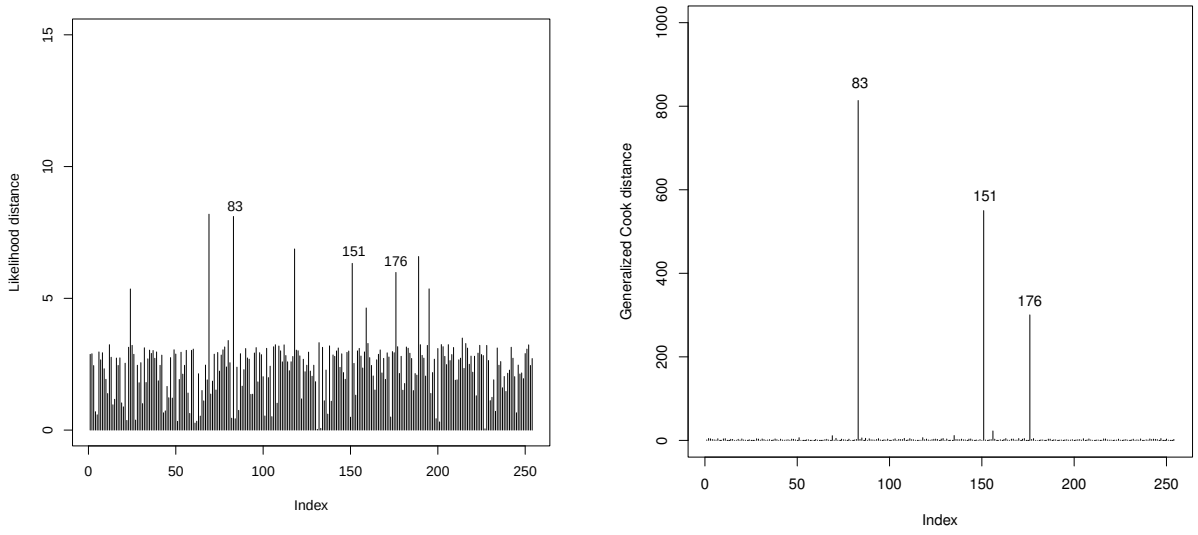
Next, utilizing the new regression model proposed, the systematic component (for $i = 1, \ldots, 254$) is considered:

$$\lambda_i = \exp\left(\beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \beta_3\, x_{i3} + \beta_4\, x_{i4} + \beta_5\, x_{i5} + \beta_6\, x_{i6}\right). \tag{3.11}$$

The quality of the fit of the GOLLL regression model is then evaluated. Figure 13 shows the LD and GCD measures for identifying possibly influential observations. They demonstrate that the 83rd, 151st and 176th observations (corresponding to the counties listed below) may be influential.

- 83rd: Gaines county with VR: 0.222, HP: 1, PR: 0.142, MS: 0, HR: 0.62, BA: 0.80 and HT: 0.063;
- 151st: Loving county with VR: 0.189, HP: 0, PR: 0.186, MS: 0, HR: 0.97, BA: 0.97 and HT: 0.05;
- 176th: Newton county with VR: 0.251, HP: 0, PR: 0.206, MS: 1, HR: 0.81, BA: 0.75 and HT: 0.105.

Gaines[8] and Loving[9] are the two counties with medium average vulnerability among the three potentially influential observations. This is explained by the extremely high health system challenges, as well as the potential impact of minorities or non-English speakers on the vaccination rate and model. The final one, Newton[10] county, has a high vulnerability level, which is explained by extremely high rates of unemployment and low income as well as significant health system challenges that may also affect the model and the immunization rate.



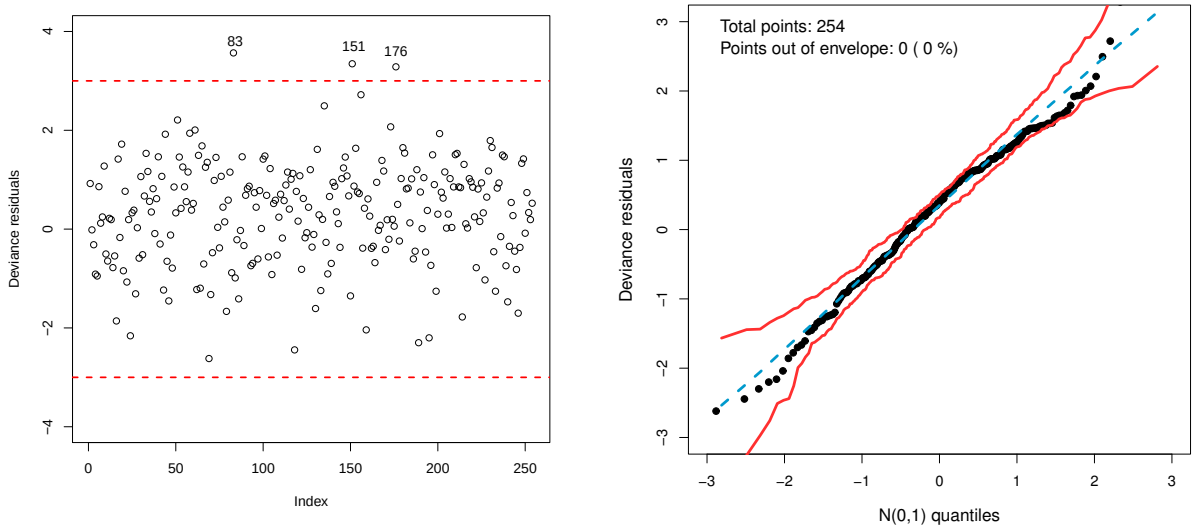
(a) (b)

Figure 13 – The GOLLL regression model. (a) LD. (b) GCD.

Furthermore, the plot of the deviance residuals in Figure 14(a) indicates that they are distributed randomly within the bands. The normal probability plot with simulated envelope in Figure 14(b) demonstrates the model's ability to fit the data set. As a result, the GOLLL regression model provides a good fit and the impacts of potentially influential observations on the regression model are insignificant.

The MLEs, SEs and p-values for the fitted GOLLL regression model to the current data are reported in Table 8.

[8] <https://covidactnow.org/us/texas-tx/county/gaines_county/?s=49703715>
[9] <https://covidactnow.org/us/texas-tx/county/loving_county/?s=49703715>
[10] <https://covidactnow.org/us/texas-tx/county/newton_county/?s=49703715>



(a) (b)

Figure 14 – The GOLLL regression model. (a) Deviance residual index. (b) Simulated envelope.

Table 8 – Fitted GOLLL regression of COVID-19 completed primary vaccination rates data.

Parameter	Estimate	SE	p-value
β_0	1.017	0.347	0.004
β_1	-0.010	0.003	0.003
β_2	-0.524	0.243	0.032
β_3	-0.149	0.031	$<$ 0.001
β_4	0.408	0.197	0.039
β_5	0.637	0.223	0.005
β_6	2.275	0.995	0.023

3.5.3 Discussion

The model checks reveal that the GOLLL regression model is suitable to explain the vaccination rates in Texas counties. From the parameter estimates reported in Table 8, the GOLLL regression model becomes:

$$\hat{\lambda}_i = \exp\left(1.017 - 0.010\,x_{i1} - 0.524\,x_{i2} - 0.149\,x_{i3} + 0.408\,x_{i4} + 0.637\,x_{i5} + 2.275\,x_{i6}\right). \quad (3.12)$$

Several facts can be drawn from Equation (3.12). For each covariate, the study reveals findings that indicate the importance of considering this model for future applications with diverse other vaccination data.

Interpretations for systematic structure λ

- All variables are statistically significant at a significance level of 5%;

- The β_1 shows a slight negative estimate. Less hospitals reporting immunization may result in a decrease in the shape parameter as a result of the skewness of the baseline distribution, which can pull the mean of vaccination rates towards lower values;

- The β_2 and β_3 of the variables PR and MS, respectively, are significant and their estimate is negative. Counties with a high poverty rate and non-metropolitan locations could lead to a decrease in shape and consequently in skewness, which shifts vaccination rates downward on average;

- The coefficients β_4, β_5, and β_6 for HR, BA, and HT have positive estimates. Counties with high rates of high school education, increased broadband access and a large percentage of heart disease patients might result in an increase in the shape of the baseline, hence skewness, altering vaccination rates to mean towards higher levels.

3.6 CONCLUDING REMARKS

The Chapter introduced the *generalized odd log-logistic Lindley* distribution and proposed the regression model with a shape systematic component to examine the factors determining COVID-19 immunization rates. Some mathematical features of this model were presented, the parameters were estimated using the maximum likelihood method and simulations were performed to evaluate them. The parameters of the suggested regression model were tested using Monte Carlo simulation, which revealed the estimators' consistency. Diagnostic analysis and deviance residuals confirmed the new model's adequacy.

A county-level investigation of COVID-19 completed primary vaccination rates in Texas, US, revealed notable findings. All of the variables were statistically significant at a level of 5%. The variables total number of hospitals reporting vaccination, percentage of individuals with income below the poverty line, and metropolitan status of the counties all had negative estimates, indicating an impact on the skewness of the baseline distribution that might push the average vaccination rates lower. Furthermore, the proportion of individuals who have completed high school, households rate that have access to broadband internet and the percentage of individuals with chronic heart disease have positive estimates, indicating an impact on the skewness that could displace the mean vaccination rates higher.

The new model showed that it was more flexible than some competitive models. Hence, it is possible to conclude that the proposed model can provide better insights into the relationship

between the explanatory variables and the response variable, serving as an alternative model to evaluate other research and improve vaccination efforts.

4 A INNOVATIVE EXPONENTIAL REGRESSION MODEL FOR EPIDEMIOLOGICAL DATA

RESUMO

A distribuição exponencial log-logística *odd* generalizada (QOSHJA; MUÇA, 2018) e (AFIFY et al., 2021) é apresentada e um modelo de regressão com um componente estrutural de forma é proposto para examinar os casos semanais de dengue no Distrito Federal do Brasil. Uma revisão das propriedades matemáticas é fornecida, o método de máxima verossimilhança é usado para estimar os parâmetros e, por meio de simulações de Monte Carlo, a precisão dos estimadores é investigada. O ajuste do modelo é avaliado usando medidas de influência global e análise de resíduos. O pressuposto de dependência temporal é relaxado para o período de análise devido ao pequeno número de observações e análises do modelo. Para o cenário temporal estudado, o modelo introduzido identificou os meses com impacto no parâmetro de forma. Finalmente, são abordadas algumas interpretações e uma discussão fornece resultados que ajudam a compreender melhor o conjunto de dados.

Palavras-chaves: Dengue. Dados epidemiológicos. Distribuição exponencial. Família log-logística *odd* generalizada. Máxima verossimilhança. Modelo de regressão. Simulação.

ABSTRACT

The generalized odd log-logistic exponential distribution (QOSHJA; MUÇA, 2018) and (AFIFY et al., 2021) is presented and a regression model with a structural shape component is proposed to examine dengue fever weekly cases in the Federal District of Brazil. A review of the mathematical properties is provided, the maximum likelihood method is used to estimate the parameters, and, through Monte Carlo simulations, the accuracy of the estimators is investigated. The model's fit is assessed using global influence metrics and residual analysis. The time dependency assumption is relaxed for the period of analysis due to the small number of observations and model analyses. For the time scenario studied, the introduced model identified months that have an impact on the shape parameter. Finally, some interpretations are addressed and a discussion provides results that help to better understand the data set.

Keywords: Dengue fever. Epidemiological data. Exponential distribution. Generalized odd log-logistic family. Maximum likelihood. Regression model. Simulation.

4.1 INTRODUCTION

Dengue fever represents a major challenge in epidemiology around the world and specially in Brazil. Recurring epidemics happen in some endemic territories in Brazil. The uncontrollable increase in the metropolitan population and the lack of information and control are factors that are determinants of the occurrence of the disease and the burden on the healthcare system. More than 100 countries have dengue fever as an endemic problem which affects millions of people each year. One of the most important factors that contributes to the global spread of the virus nowadays is the climate change that is growing up fast.

In this regard, many applications for generalized extreme value (GEV) distributions can be found in epidemiological studies. Li et al. (2020) presented a review study on the relationship between dengue fever and meteorological parameters, as well as a meta-analysis to investigate the impact of ambient temperature and precipitation on dengue fever. Dengue cases counts during outbreaks in Thailand were modeled using extreme value theory (EVT) (LIM; DICKENS; COOK, 2020). A zero-inflated GEV regression model is used in an application using Vietnam dengue data (DIOP; DEME; DIOP, 2021). The estimate of the risk of infection for the individuals is based on the covariates of age and weight.

Likewise, Marani et al. (2021) presented a GEV approach to study the frequency and intensity of extreme novel epidemics similar to COVID-19. As well, an extreme value statistics to predict in real-time severe influenza epidemics is used (THOMAS; ROOTZÉN, 2022). Moreover, a study investigate the extreme correlation between infectious illness outbreaks and crude oil futures (LIN; ZHANG, 2022). Tian et al. (2022) estimated the disease burden of dengue in endemic regions to study the influence pattern of socioeconomic factors. They recommend allocating more resources to areas with high population expansion and urbanization.

Further, Lun et al. (2022) explored the characteristics and studied the temporal-spatial cases of the overseas imported dengue fever in outbreak provinces of China. Sandeep et al. (2023) studied myocarditis manifestations in dengue cases in a systematic review. In Oliveira-Júnior et al. (2023), dengue fever cases in the Brazilian state of Alagoas were modeled monthly using GEV distribution. The findings underline the significance of ongoing monitoring and assistance in this area.

Therefore, the study is based on the dengue fever weekly cases of the epidemiological weeks in 2022 in the Federal District of Brazil. A regression analysis is applied to study extreme events (an epidemiological event that affects health centers and the economy) and

the maximum likelihood method is used to estimate the parameters. The accuracy of the estimators is confirmed by means of Monte Carlo simulations. Some local influence measures are used, as are residual analyses, to validate the goodness-of-fitness of the proposed model.

Considering this, the Chapter presents a new regression model based on the *generalized odd log-logistic exponential* (GOLLE) distribution (QOSHJA; MUÇA, 2018) and (AFIFY et al., 2021). The GOLL-G class's broad flexibility, which allows modeling its tails and assymetry, combined with the exponential distribution, which is used in many EVT applications and has a closed mathematical form, makes the novel regression a relevant model to apply in many areas, as well as a potential application in extreme events.

This Chapter is organized as follows. Section 4.2 presents the GOLLE distribution (QOSHJA; MUÇA, 2018) and (AFIFY et al., 2021). Section 4.3 exhibits the linear representation and some of its mathematical features. Maximum likelihood estimation is discussed, as well as a Monte Carlo simulation study to show the estimators' consistency. In Section 4.4, a novel regression model based on the GOLLE distribution, associating the covariates with the baseline shape parameter, is introduced and simulations examine the behavior of the estimators. Section 4.5 demonstrates the usefulness of the new regression model using an epidemiological data set and discusses some findings. Additionally, diagnostic and residual analyses are provided. Section 4.6 contains some final observations.

4.2 THE GOLLE DISTRIBUTION

The cdf and the pdf, respectively, describe the exponential distribution (for $x > 0$), as:

$$G(x;\lambda) = 1 - \mathrm{e}^{-\lambda x} \tag{4.1}$$

and

$$g(x;\lambda) = \lambda \mathrm{e}^{-\lambda x}, \tag{4.2}$$

for shape parameter $\lambda > 0$.

The cdf and the pdf of the GOLLE distribution (QOSHJA; MUÇA, 2018) and (AFIFY et al., 2021) are defined, respectively, as:

$$F(y;\alpha,\theta,\lambda) = \frac{\left(1-\mathrm{e}^{-\lambda y}\right)^{\alpha\theta}}{\left(1-\mathrm{e}^{-\lambda y}\right)^{\alpha\theta} + \left[1-\left(1-\mathrm{e}^{-\lambda y}\right)^{\theta}\right]^{\alpha}} \tag{4.3}$$

and

$$f(y;\alpha,\theta,\lambda) = \frac{\alpha\,\theta\,\lambda\,\mathrm{e}^{-\lambda y}\,\left(1-\mathrm{e}^{-\lambda y}\right)^{\alpha\theta-1}\,\left[1-\left(1-\mathrm{e}^{-\lambda y}\right)^{\theta}\right]^{\alpha-1}}{\left\{\left(1-\mathrm{e}^{-\lambda y}\right)^{\alpha\theta}+\left[1-\left(1-\mathrm{e}^{-\lambda y}\right)^{\theta}\right]^{\alpha}\right\}^{2}}. \tag{4.4}$$

The hrf can be obtained by substituting Equations (3.2) and (3.1) into Equation 2.4, as:

$$\tau(y;\alpha,\theta,\lambda) = \frac{\alpha\theta\left(\lambda\mathrm{e}^{-\lambda x}\right)\left(1-\mathrm{e}^{-\lambda x}\right)^{\alpha\theta-1}}{\left[1-\left(1-\mathrm{e}^{-\lambda x}\right)^{\theta}\right]\left\{\left(1-\mathrm{e}^{-\lambda x}\right)^{\alpha\theta}+\left[1-\left(1-\mathrm{e}^{-\lambda x}\right)^{\theta}\right]^{\alpha}\right\}}. \tag{4.5}$$

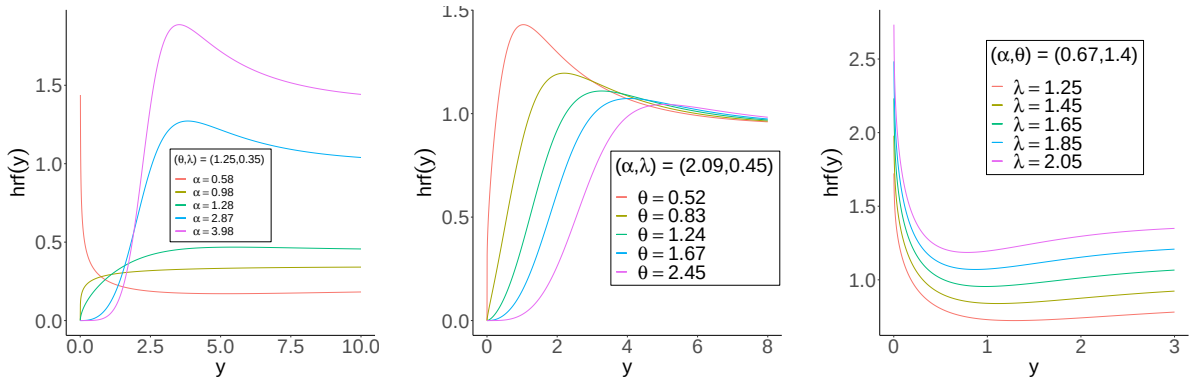
Here in the Table 9, the sub-models derived from Equation (4.4) are presented. Its ability to handle data fitting across a variety of distributions highlights its adaptability and applicability.

Table 9 – Submodels associated to the GOLLE distribution.

α	θ	λ	**Submodel**
-	1	-	Odd log-logistic exponential (OLLE) distribution (GLEATON; LYNCH, 2006)
1	-	-	Exponentiated-exponential (Exp-E) distribution (GUPTA; KUNDU, 2001)
1	1	-	Exponential (E) distribution

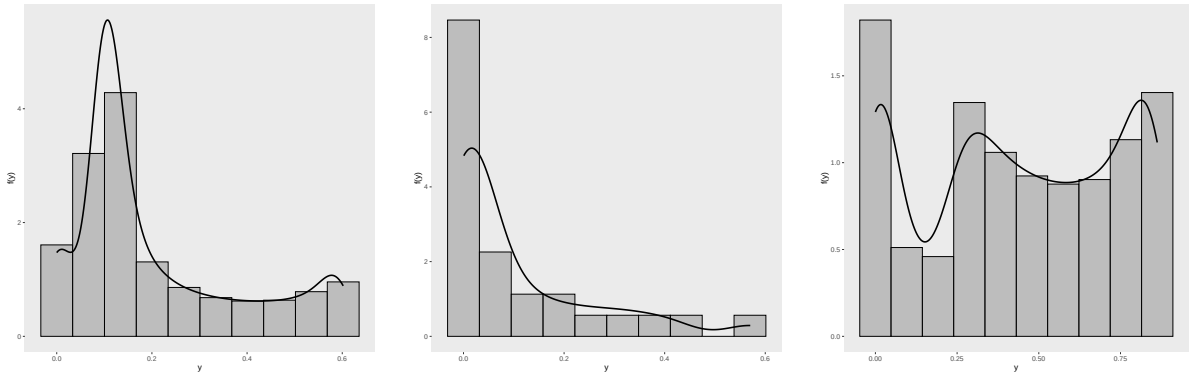
Plots of the hrf and the histograms of Y for chosen parameters are shown in Figures 15 and 16. The versatility of the GOLLE distribution in generating a wide range of hazard shapes is one of its most notable features compared to the constant behavior over time of the exponential hrf.

The inverse J-shape, increasing-decreasing, decreasing-increasing, bathtub and several more shapes are shown in Figure 15. In comparison to other class of distributions, Figure 16, showed the model as a powerful tool for modeling non-normal data sets with a wide range of histogram patterns (asymmetric, heavy tail, multimodal, etc).



(a) (b) (c)

Figure 15 – GOLLE hrf for selected values.



(a) (b) (c)

Figure 16 – GOLLE histogram. (a) GOLLE(0.15,73,2.50). (b) GOLLE(0.22,1.13,7.50). (c) GOLLE(0.07,120.13,3.50).

4.3 MAIN PROPERTIES

This Section showed the GOLLE distribution's linear representation of the density function, the qf, moments and mgf, as described in (AFIFY et al., 2021).

Definition 1. *The GOLLE density (4.4) can be represented linearly using exponential densities, as:*

$$f(y;\alpha,\theta,\lambda) = \sum_{k,m=0}^{\infty} t_{k,m}\, g(y;\lambda^*), \tag{4.6}$$

with a shared parameter $\lambda^* = \lambda^*(\lambda, m) = \lambda(m+1)$ *and* $t_{k,m}$ *defined by the quantities below:*

$$t_{k,m} = \frac{(-1)^m \binom{k}{m}}{(m+1)} b_k.$$

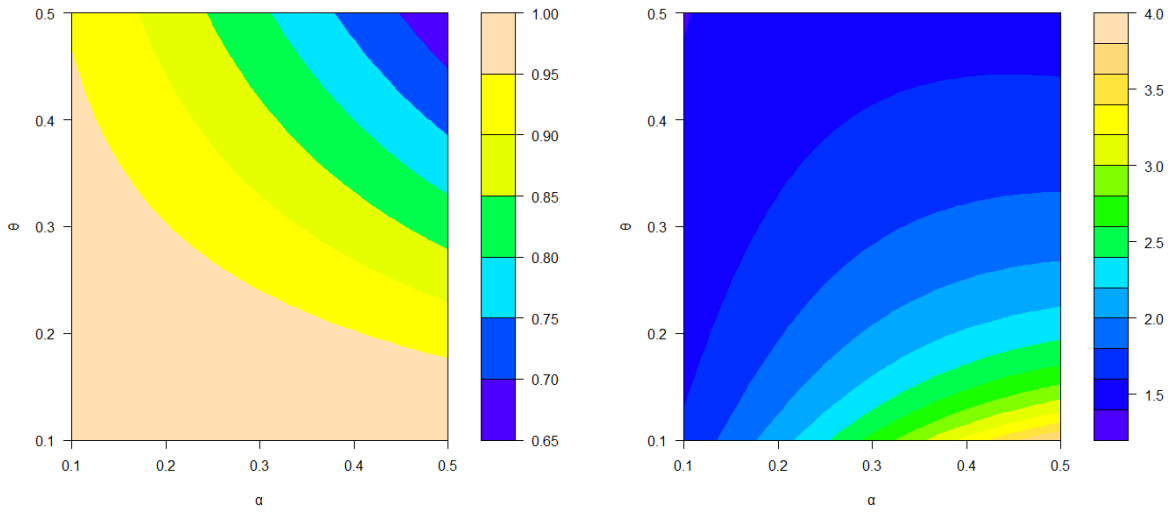
The Definition (1) indicates that numerous mathematical properties of the GOLLE distribution can be obtained directly from those of the exponential distribution.

4.3.1 Quantile function

The qf of Y is simply found, as:

$$Q(u) = -\frac{1}{\lambda}\log[1-\varepsilon_{\alpha.\theta}(u)]. \tag{4.7}$$

Figure 17 displays Galton's skewness and Moors' kurtosis varying α and θ, with $\lambda = 1.58$. The influence of both parameters on the distribution shape is shown in these plots. As the parameters α and θ increase, the skewness and kurtosis measures decrease to a minimal value.



(a) (b)

Figure 17 – GOLLE distribution. (a) Galton's skewness. (b) Moors' kurtosis.

4.3.2 Moments

The moments of the GOLLE distribution are presented in (AFIFY et al., 2021).

Definition 2. *The nth moment of the GOLLE distribution is defined by:*

$$\mu'_n = \mathbb{E}(Y^n) = \sum_{k,m=0}^{\infty} \frac{n!}{\lambda^{*n}} t_{k,m} = \sum_{k,m=0}^{\infty} \frac{(-1)^m \binom{k}{m}}{[(m+1)\lambda^*]^n} b_k.$$

4.3.3 Generation function

The gf of the GOLLE distribution is shown in (AFIFY et al., 2021).

Definition 3. *The mgf of the GOLLE density can be expressed, as:*

$$M_Y(t) = \sum_{k,m=0}^{\infty} \frac{\lambda^*}{\lambda^* - t} t_{k,m} = \sum_{k,m=0}^{\infty} \frac{(-1)^m \binom{k}{m} \lambda^*}{\lambda^* - t} b_k, \quad \textit{for} \quad t < \lambda^*.$$

4.3.4 Estimation

The MLEs of the GOLLE parameters vector $\boldsymbol{\psi} = (\alpha, \theta, \lambda)^\top$ are calculated from a complete sample $y_1, \ldots, y_n$ by maximizing:

$$l_n(\boldsymbol{\psi}) = n\log(\alpha\theta\lambda) - \lambda\sum_{i=1}^{n} y_i + (\alpha\theta - 1)\sum_{i=1}^{n}\log(1 - \mathrm{e}^{-\lambda y_i}) + (\alpha - 1)\sum_{i=1}^{n}\log\left[1 - (1 - \mathrm{e}^{-\lambda y_i})^{\theta}\right] - 2\sum_{i=1}^{n}\log\left\{(1 - \mathrm{e}^{-\lambda y_i})^{\alpha\theta} + \left[1 - (1 - \mathrm{e}^{-\lambda y_i})^{\theta}\right]^{\alpha}\right\}. \tag{4.8}$$

Let's consider

$$A_i(\lambda) = A_i = 1 - \mathrm{e}^{\lambda y_i}.$$

Therefore, the elements of the score vector can be formulated, as follows:

$$\begin{aligned} U_\alpha &= \frac{n}{\alpha} + \theta\sum_{i=1}^{n}\log(A_i) + \sum_{i=1}^{n}\log\left(1 - {A_i}^{\theta}\right) \\ &- 2\sum_{i=1}^{n}\frac{\theta\log(A_i){A_i}^{\alpha\theta} + (1 - {A_i}^{\theta})^{\alpha}\log(1 - {A_i}^{\theta})}{{A_i}^{\alpha\theta} + (1 - {A_i}^{\theta})^{\alpha}}, \end{aligned}$$

$$\begin{aligned} U_\theta &= \frac{n}{\theta} + \alpha\sum_{i=1}^{n}\log(A_i) - (\alpha - 1)\sum_{i=1}^{n}\frac{{A_i}^{\theta}\log(A_i)}{1 - {A_i}^{\theta}} \\ &+ \sum_{i=1}^{n}\frac{\alpha{A_i}^{\alpha\theta}\log(A_i) + (1 - {A_i}^{\alpha})^{\theta}\log(1 - {A_i}^{\alpha})}{{A_i}^{\alpha\theta} + (1 - {A_i}^{\alpha})^{\theta}} \end{aligned}$$

and

$$\begin{aligned} U_\lambda &= \frac{n}{\lambda} - \sum_{i=1}^{n} y_i + (\alpha\theta - 1)\sum_{i=1}^{n}\frac{(1 - A_i)}{A_i} - \theta(\alpha - 1)\sum_{i=1}^{n}\frac{(1 - A_i)A_i^{\theta-1}}{1 - A_i^{\theta}} \\ &- 2\alpha\theta\sum_{i=1}^{n}\frac{(1 - A_i)[A_i^{\alpha\theta-1} - A_i^{\theta-1}(1 - A_i^{\theta})^{\alpha-1}]}{A_i^{\alpha\theta} + (1 - A_i^{\theta})^{\alpha}}. \end{aligned}$$

Using a Newton-Raphson type method and setting the score equations $U_\alpha = U_\theta = U_\lambda = 0$, the MLEs are calculated. The *optim* procedure can also be used to numerically maximize Equation (4.8).

4.3.5 Simulation study

In two scenarios, Monte Carlo simulations generated by $1,000$ samples of varied sizes of the GOLLE distribution are utilized to assess the accuracy of MLEs. For each sample size, $n = \{50, 100, 200, 400, 800,\ 1,000\}$, the AEs, ABs and RMSEs are computed, for each $\boldsymbol{\epsilon} = (\alpha, \theta, \lambda)$.

Table 10 – Simulations results for GOLLE distribution

scenario 1 - GOLLE(0.23,1.25,0.89)									
Par	n=50			n =150			n = 300		
	AE	AB	RMSE	AE	AB	RMSE	AE	AB	RMSE
α	0.274	0.044	0.207	0.266	0.036	0.141	0.249	0.019	0.088
θ	1.580	0.330	0.909	1.357	0.107	0.618	1.292	0.042	0.417
λ	1.076	0.186	0.603	0.932	0.042	0.399	0.908	0.0018	0.281
Par	n = 500			n = 750			n = 1000		
α	0.239	0.009	0.060	0.234	0.008	0.041	0.235	0.005	0.034
θ	1.279	0.029	0.309	1.244	0.006	0.211	1.252	0.002	0.179
λ	0.906	0.016	0.213	0.884	0.006	0.147	0.890	0.000	0.124
scenario 2 - GOLLE(0.85,0.15,1.15)									
Par	n = 50			n = 150			n = 300		
α	0.824	0.026	0.292	0.837	0.012	0.184	0.855	0.005	0.121
θ	0.201	0.051	0.175	0.164	0.014	0.051	0.154	0.004	0.024
λ	2.232	1.082	3.356	1.479	0.329	0.980	1.239	0.089	0.446
Par	n = 500			n = 750			n = 1000		
α	0.848	0.002	0.090	0.851	0.001	0.062	0.849	0.001	0.054
θ	0.153	0.003	0.017	0.151	0.001	0.011	0.151	0.001	0.010
λ	1.225	0.075	0.338	1.172	0.022	0.218	1.176	0.026	0.196

As predicted by the consistency requirement, the results in Table 10 indicate that AEs approximate to the real values and ABs and RMSE approach zero as n increases. It is notable that for scenario 1, all the estimates obtained when $n = 50$ were overestimated, while for scenario 2, the parameters θ and λ were overestimated. This shows the sensitivity of the model's parameters to some values, but in general, as the sample size increases, convergence towards the true values is achieved.

4.4 THE GOLLE REGRESSION MODEL

The systematic component of the GOLLE regression model takes into account the fact that the shape parameter λ in Equation (4.4) varies between observations (for $i = 1, \ldots, n$), as:

$$\lambda_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \tag{4.9}$$

where is a twice continuously differentiable log-linear link function and $\boldsymbol{\beta} = (\beta, \ldots, \beta_p)^\top$ is the parameter vector of dimension p associated with the explanatory variables $\boldsymbol{x_i}^\top = (x_{i1}, \ldots, x_{ip})$.

4.4.1 Estimation

Except for λ_i, the components of the score vector for U_α and U_θ are the same as those obtained from the Equations presented in Subsection (4.3.4). The score component of the vector parameter λ_i is defined to add the regression part in the manner described below:

$$\begin{aligned} U_{\lambda_i} &= \sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta}} g(x_i;\lambda_i)}{g(x_i;\lambda_i)} + (\alpha\theta - 1)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta}} G(x_i;\lambda_i)}{G(x_i;\lambda_i)} + \theta(1-\alpha)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta}} G(x_i;\lambda_i) G(x_i;\lambda_i)^{\theta-1}}{1 - G(x_i;\lambda_i)^{\theta}} \\ &- 2\sum_{i=1}^{\infty} \partial_{\boldsymbol{\beta}} G(x_i;\lambda_i) \frac{G(x_i;\lambda_i)^{\alpha\theta-1} - G(x_i;\lambda_i)^{\theta-1}[1 - G(x_i;\lambda_i)^{\theta}]^{\alpha-1}}{G(x_i;\lambda_i)^{\alpha\theta} + [1 - G(x_i;\lambda_i)^{\theta}]^{\alpha}}, \end{aligned}$$

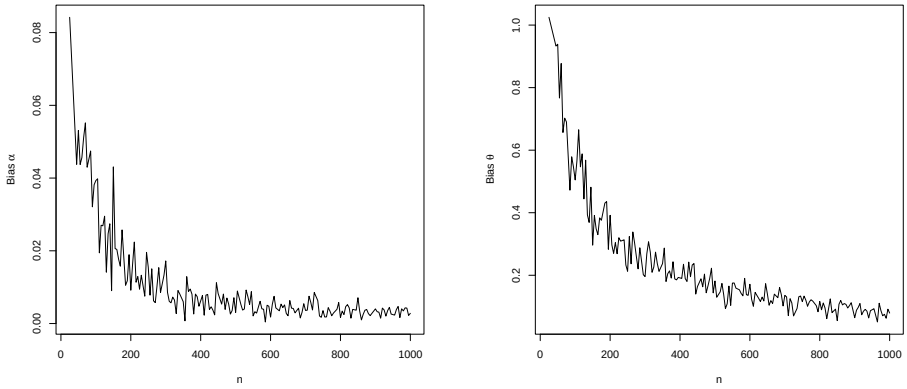
where $\partial_{\boldsymbol{\beta}} g(x_i;\lambda_i) = \partial_{\lambda_i} g(x_i;\lambda_i)\partial_{\boldsymbol{\beta}}\lambda_i(x_i;\boldsymbol{\beta})$ and $\partial_{\boldsymbol{\beta}} G(x_i;\lambda_i) = \partial_{\lambda_i} G(x_i;\lambda_i)\partial_{\boldsymbol{\beta}}\lambda_i(x_i;\boldsymbol{\beta})$ denotes the derivatives of the parameter λ_i using the chain rule.

The MLE $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ of the regression model is calculated setting the score equations $U_\alpha = U_\theta = U_{\lambda_i} = 0$ using an iterative method algorithm to find roots or using the *optim* routine.

4.4.2 Simulation study

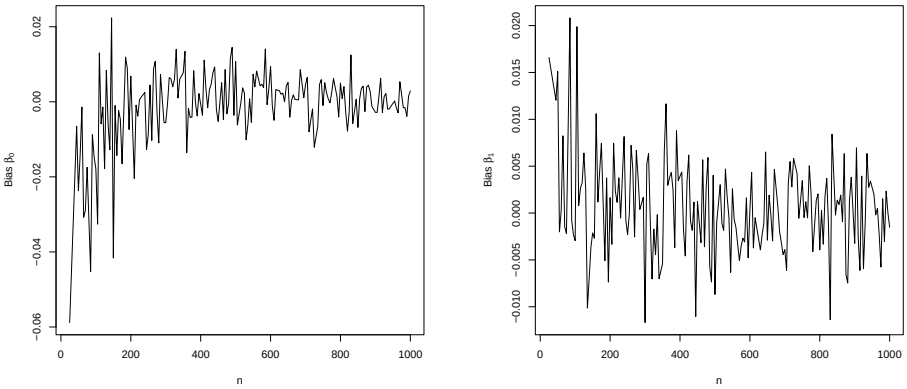
To show the accuracy of the MLEs for $\alpha = 0.75$, $\theta = 1.50$, $b = 0.50$, $\beta_0 = 0.85$ and $\beta_1 = 1.20$, 1,000 samples of size $n = \{25, \ldots, 1,000\}$ from Equation (4.7) were generated. The study is based on the measurements: biases, MSEs, ALs and CPs.

Figures 18-22 show the values of these measures in relation to n. Biases, MSEs and ALs tend to zero as sample size increases. In the biases, the estimate of β_0 is an underestimated, while the estimate of β_1 shows oscillatory behavior. These indicated potential concerns with optimization for certain values and sample sizes. Despite, the biases converges to zero as expected. Furthermore, in Figure 24, the CP is near 0.95. These findings corroborate the consistency of the MLEs.



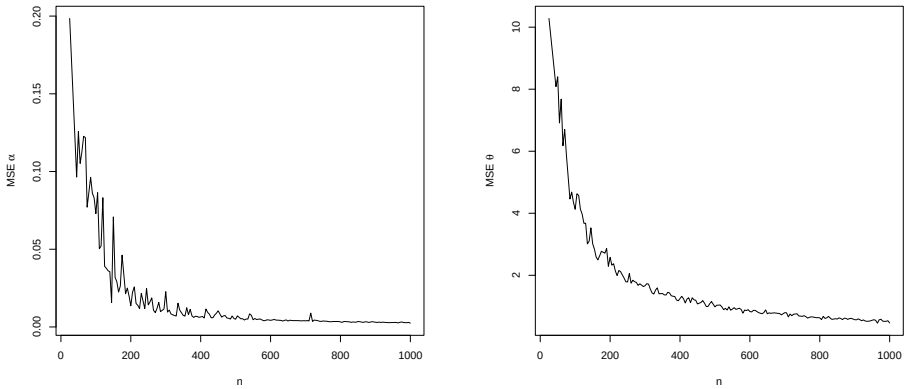
(a) (b)

Figure 18 – Biases versus sample size from GOLLE regression model.



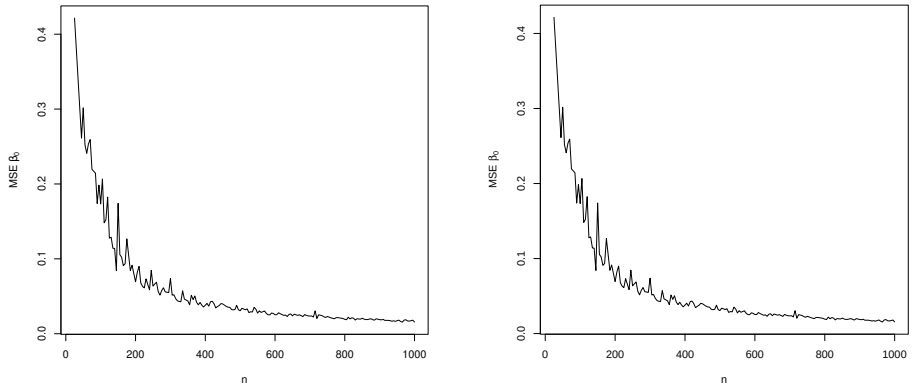
(a) (b)

Figure 19 – Biases versus sample size from GOLLE regression model.



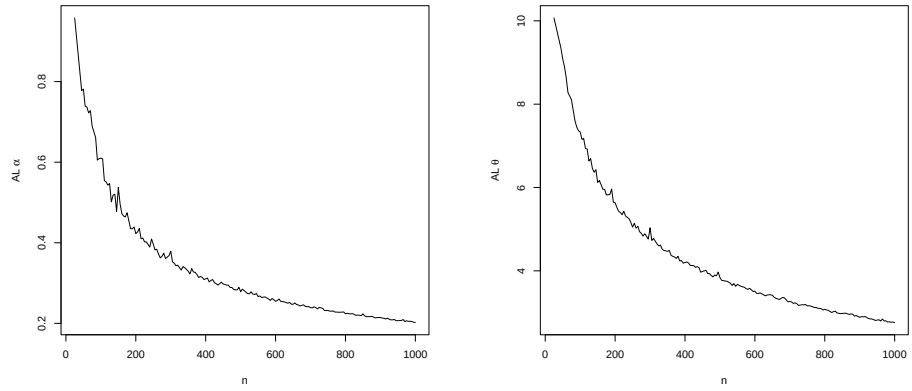
(a) (b)

Figure 20 – MSEs versus sample size from GOLLE regression model.



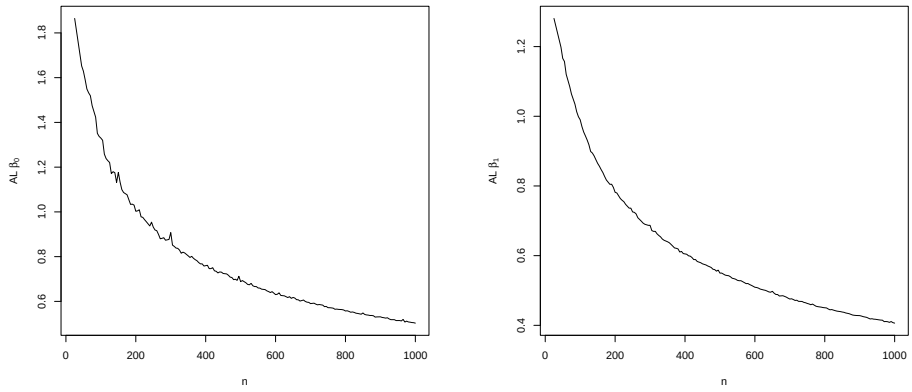
(a) (b)

Figure 21 – MSEs versus sample size from GOLLE regression model.



(a) (b)

Figure 22 – ALs versus sample size from GOLLE regression model.



(a) (b)

Figure 23 – ALs versus sample size from GOLLE regression model.

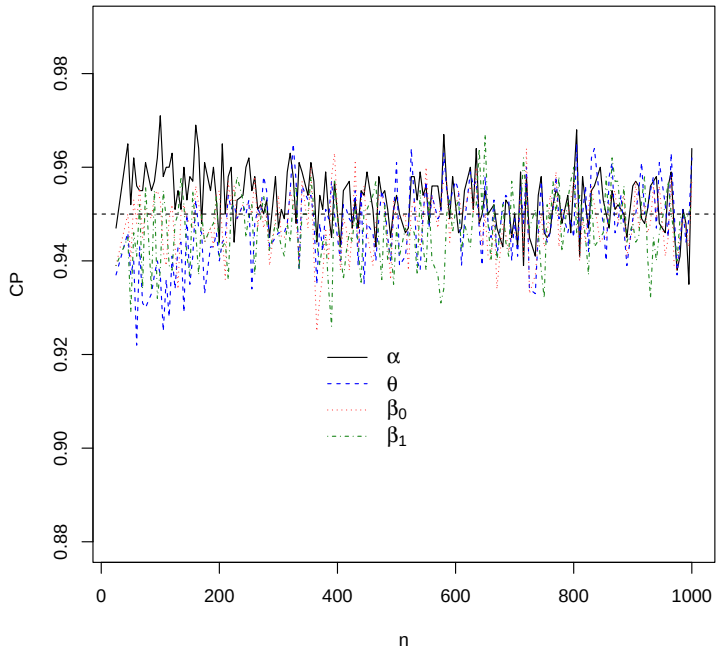


Figure 24 – CPs versus sample size from GOLLE regression model.

4.5 APPLICATION: FEDERAL DISTRICT DENGUE FEVER DATA

Table 11 illustrates some alternative distributions of some well-known generators, in addition to the nested models, that were utilized to compare with the GOLLE model.

Table 11 – Competitive distributions compared to the GOLLE distribution.

Distribution	**Reference**
Kumaraswamy-Fréchet (KwFr)	(MEAD, 2014)
Kumaraswamy -Exponential (KwE)	(ADEPOJU; CHUKWU, 2015)
Gamma-Fréchet (GFr)	(-)
Gamma-Exponentital (GE)	(KUDRIAVTSEV, 2019)
Beta-Exponentital (BE)	(NADARAJAH; KOTZ, 2006a)
Fréchet (Fr)	(FRÉCHET, 1927)

The distributions are presented (for $x > 0$), respectively, as:

$$F_{\mathsf{KwFr}}(x) = \{1 - [F_{\mathsf{Fr}}(x)]^a\}^b,$$

$$F_{\mathsf{KwE}}(x) = \{1 - [G(x)]^a\}^b,$$

$$F_{\mathsf{GFr}}(x) = \frac{\gamma\{a, -\log[1 - F_{\mathsf{Fr}}(x)]/b\}}{\Gamma(a)},$$

$$F_{\mathsf{GE}}(x) = \frac{\gamma\{a, -\log[1 - G(x)]/b\}}{\Gamma(a)},$$

$$F_{\mathsf{BE}}(x) = I_{G(x)}(a, b) = \frac{1}{B(a, b)} \int_0^{G(x)} w^{a-1}(1 - w)^{b-1} dw$$

and

$$F_{\mathsf{Fr}}(x) = \exp[-(x - a)^{-b}],$$

where all of the parameters are positive and $G(x)$ and $F_{\mathsf{Fr}}(x)$ represent the Equation (4.2) and Fréchet distributions, respectively. The *goodness.fit* function of *AdequacyModel* package computes the MLEs (SEs in parenthesis) for all fitted models using the BFGS approach.

4.5.1 A data set definition

The data set was obtained from the Federal District Health Department (SES-DF[11]). This system maintains a register of patient notifications that include diseases, injuries and public health incidents that are required to be reported. This contains epidemiological information of dengue fever, zika fever, chikungunya fever, yellow fever and more than 40 other diseases. The data is made up of notifications relating to dengue fever cases (in thousands) registered in the Federal District of Brazil and it spans all 49 epidemiological weeks (observations) in 2022:

- y_i: total dengue fever cases (in thousands) of a epidemiological week (DG) (response variable);
- m_{ij}: month (levels: 0 - January to 11 - December). Thus, for $i = 1, \ldots, 49$ and $j = 0, \ldots, 11$, dummy variables.

The proposed model has both advantages and disadvantages over counting models. The exponential distribution utilized as a baseline has numerous notable features, including a usefulness in some epidemiological cases, memorylessness, a good fit with empirical data with heavy tails, flexibility and a simple density form. On the other hand, various issues can occur, such as a lack of flexibility for trend modeling, violations of the assumption of independence,

[11] <https://www.saude.df.gov.br/informes-dengue-chikungunya-zika-febre-amarela>

which can result in erroneous models and limitations with inflated zero data, among others. Nevertheless, the model captures the significance of the exploratory variables and an extreme event of dengue fever cases in the time scenario.

Table 12 offers some descriptive statistics. The number of dengue fever cases fluctuated from very low (0.277) to high (6.726). The standard deviation is 1.445, indicating more variability in dengue fever cases over time. The distribution is skewed to the right (1.509), indicating that there are more extreme values near the top of the scale, while kurtosis indicates heavier tails (4.997).

Table 12 – Descriptive statistics of dengue fever cases.

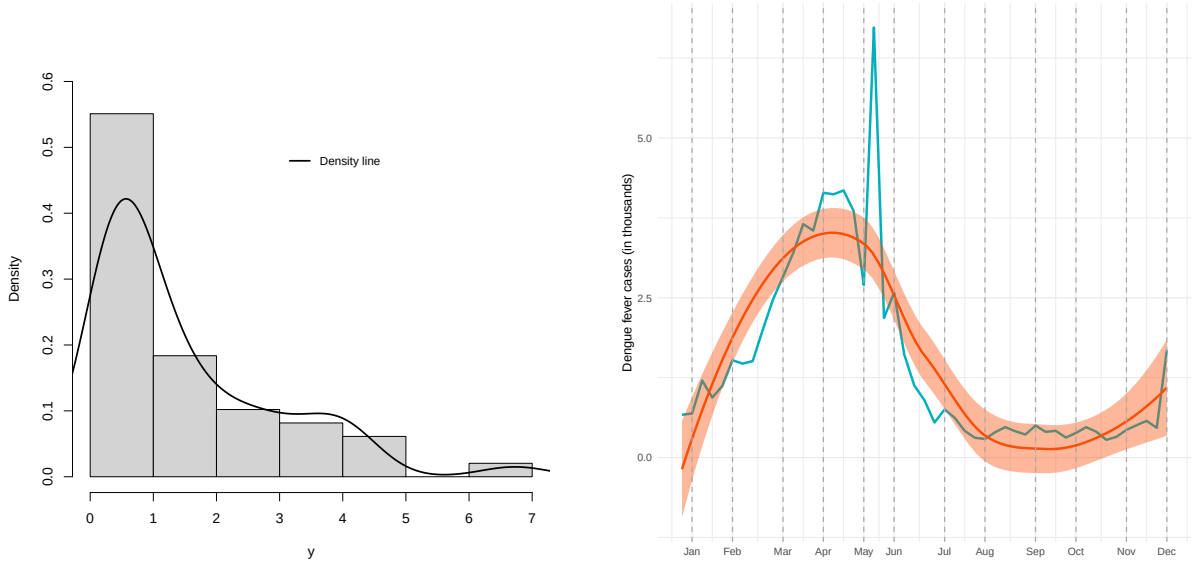
Variable	Min.	Max.	Mean	Median	SD	Skewness	Kurtosis
DG	0.277	6.726	1.483	0.752	1.445	1.509	4.997

Figure 25 illustrates the histogram and time series of the data. Figure 25(a) showed a fat tail behavior, which corroborates with extreme event data. Figure 25(b) illustrates the existence of extreme occurrences with comparable behavior during the months of May and June. The abrupt increases are a record for the same period since 1998[12] demonstrating the atypical behavior of the observations, which deviate significantly from the historical average, indicating an unusual outbreak, or, in epidemiology, an extreme event for dengue cases that can have an impact on both the health system and the economy. In addition, the plot reveals an upward trend during the months of February and June, when dengue fever is most likely to occur in the Federal District.

Brazil is a continental country and has different patterns, nevertheless, the Midwest is where there is the highest incidence of dengue fever, according to the arbovirus monitoring panel of the Ministry of Health[13], which is the region of the study data.

[12] <https://www.correiobraziliense.com.br/cidades-df/2022/06/5017446-casos-atingem-maior-numero-desde-1998.html>

[13] <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/aedes-aegypti/monitoramento-das-arboviroses>



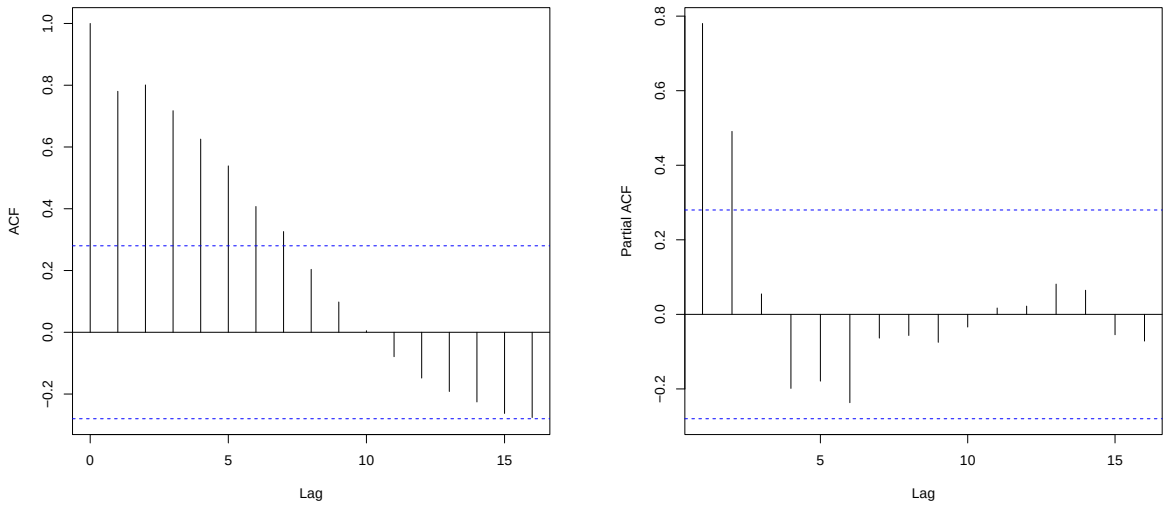
(a) (b)

Figure 25 – Dengue fever data. (a) Histogram and empirical density. (b) Variation across months with trend smoothed line.

4.5.2 Results

The study of time series data on dengue fever cases requires a thorough examination, which involves determining correlations between subsequent observations. Failure to account for these associations may result in an inaccurate model that ignores temporal relationships, potentially leading to incorrect forecasts and interpretations. As a result, it is critical to examine the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots to determine the presence of serial correlation. The ACF and PACF plots in Figure 26 indicate an autoregressive integrated moving average model with a one-lag in the differenced series (ARIMA(1,1,0)).

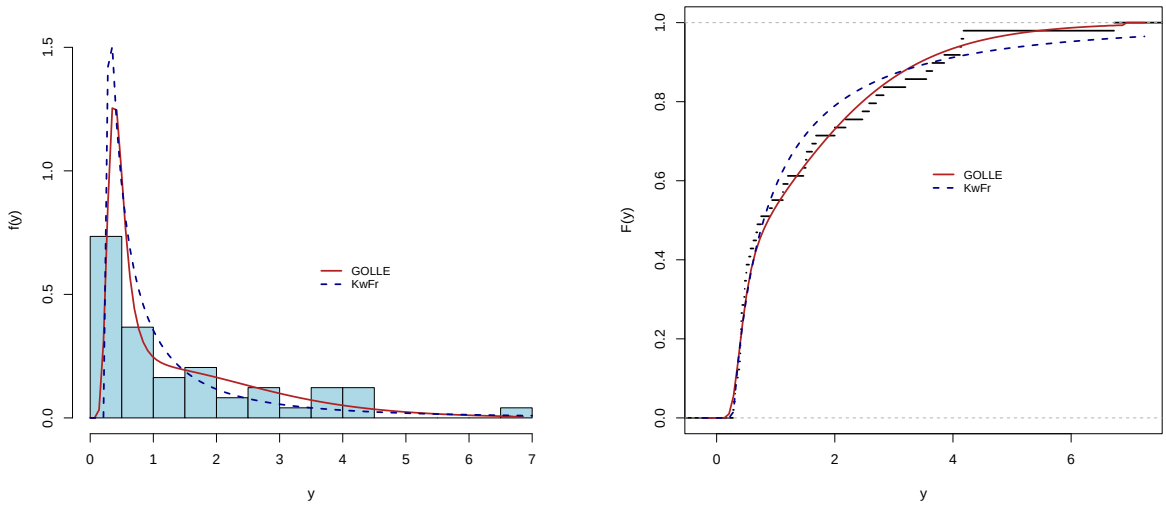
Despite the fact that dengue cases exhibit a correlation, i.e. dependence, the proposed regression model can be used provided this assumption is relaxed, given the small number of observations and the model's fit is assessed using diagnostic and residual analysis.



(a) (b)

Figure 26 – Dengue fever data. (a) ACF. (b) PACF.

Table 13 summarizes the results of the fitted distributions to the current data and demonstrates that the GOLLE distribution is the best fit. In fact, the histogram and plots of the predicted density functions in Figure 27(a) and the empirical cdf computed ones in Figure 27(b) confirm this result. The Fréchet distribution density is widely used for modeling extreme occurrences and these results show that the KwFr distribution (the second best model) is competitive with the presented model.



(a) (b)

Figure 27 – Fitted models of dengue fever data. (a) Histogram and estimated pdfs. (b) Empirical and estimated cdfs.

Table 13 – Findings from the fitted models of dengue fever data.

Model	Parameters				W*	A*	KS
GOLLE(α, θ, λ)	**0.154**	**76.500**	**5.402**		**0.060**	**0.400**	**0.077**
	(0.018)	**(0.019)**	**(0.003)**				**(0.914)**
OLLE(α, λ)	1.180	1	0.634		0.318	1.929	0.160
	(0.142)	(-)	(0.086)				(0.145)
EE(θ, λ)	1	1.391	0.830		0.313	1.898	0.175
	(-)	(0.284)	(0.147)				(0.088)
E(λ)	1	1	0.674		0.316	1.913	0.170
	(-)	(-)	(0.096)				(0.103)
KwFr(β, γ, a, b)	3.851	51.070	0.172	0.271	0.087	0.559	0.097
	(1.409)	(71.389)	(0.060)	(0.008)			(0.705)
KwE(β, γ, λ)	4.500	0.151	5.402		0.242	1.482	0.205
	(0.005)	(0.022)	(0.003)				(0.028)
GFr(β, a, b)	0.465	0.777	0.225		0.128	0.830	0.120
	(0.082)	(0.142)	(0.039)				(0.443)
BE(β, γ, λ)	3.027	0.150	5.402		0.253	1.548	0.197
	(0.1.054)	(0.023)	(0.003)				(0.038)
GE(β, λ)	1.323	0.892			0.317	1.917	0.173
	(0.241)	(0.197)					(0.096)
Fr(a,b)	1.791	-0.281			0.235	1.449	0.173
	(0.285)	(0.076)					(0.094)

LR tests were used to compare the GOLLE distribution and its nested models. Table 14 shows that adding more parameters has a significant influence on accurately modeling the existing data.

Table 14 – LR tests of the GOLLE distribution

Models	**Statistic w**	**p-value**
GOLLE vs E	29.657	< 0.0001
GOLLE vs EE	27.143	< 0.0001
GOLLE vs OLLE	27.937	< 0.0001

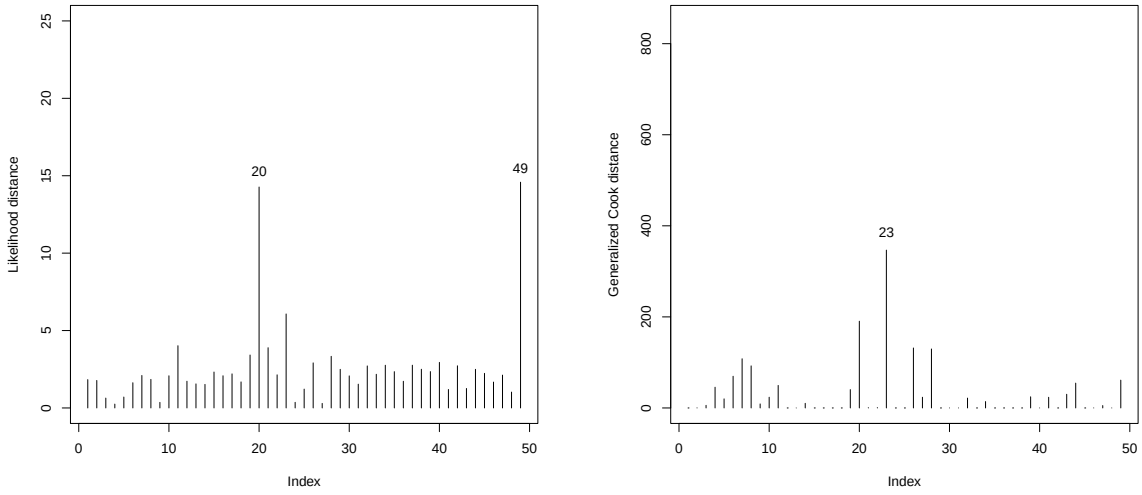
The systematic structures are considered here based on a non-linear equation (for $i = 1, \ldots, 49$), as follows:

$$\lambda_i = \exp\left(\beta_0 + \sum_{j=1}^{11} \beta_j \, m_j\right).$$

One influential observation was indicated using the GCD and LD measures, as shown in Figure 28. It's worth noting that the 49th observation (corresponding to the last epidemiological week) has the potential to be influential. One probable explanation, as an important event, is the vacation/recess period at the end of the year, which causes a backlog of alerts

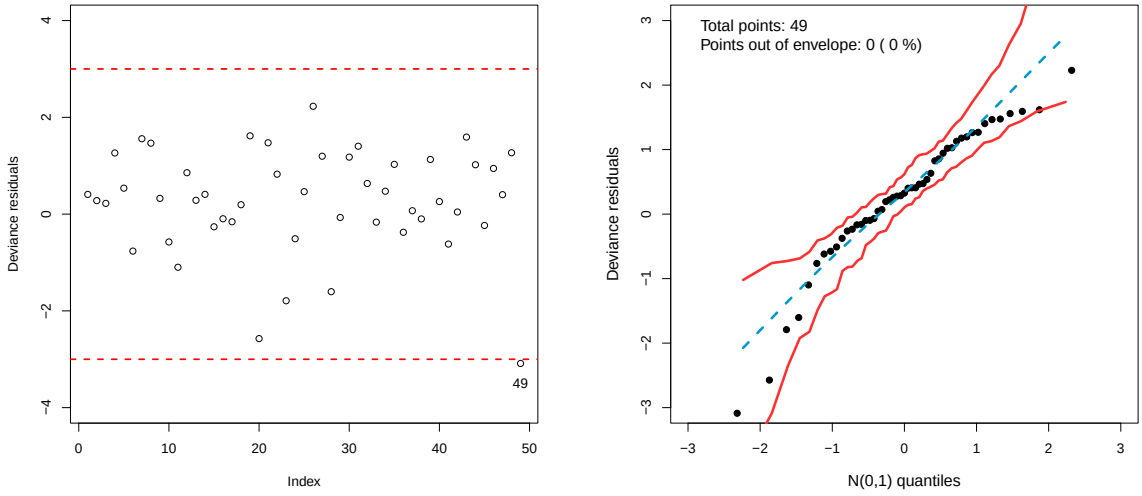
due to a lack of health care personnel in service who notify cases and input data into the system.

Nonetheless, Figure 29 indicates that the index deviation residuals behave randomly over the range and the residuals are within the simulated envelope, showing that the observation has no significant impact on the regression model.



(a) (b)

Figure 28 – The GOLLE regression model. (a) LD. (b) GCD.



(a) (b)

Figure 29 – The GOLLE regression model. (a) Deviance residual index. (b) Simulated envelope.

The results (MLEs, SEs, and p-values) of the fitted GOLLE regression model to the current

data are presented in Table 15.

Table 15 – Fitted GOLLE regression of dengue fever data.

Parameter	Estimate	SE	p-value
$\boldsymbol{\beta_0}$	$\mathbf{-9.1118}$	$\mathbf{2.3349}$	$\mathbf{0.0004}$
$\boldsymbol{\beta_1}$	$\mathbf{0.5459}$	$\mathbf{0.1892}$	$\mathbf{0.0066}$
$\boldsymbol{\beta_2}$	$\mathbf{1.0759}$	$\mathbf{0.1928}$	$\mathbf{< 0.0001}$
$\boldsymbol{\beta_3}$	$\mathbf{1.6690}$	$\mathbf{0.1829}$	$\mathbf{< 0.0001}$
$\boldsymbol{\beta_4}$	$\mathbf{1.7019}$	$\mathbf{0.1881}$	$\mathbf{< 0.0001}$
$\boldsymbol{\beta_5}$	$\mathbf{1.4481}$	$\mathbf{0.2350}$	$\mathbf{< 0.0001}$
β_6	0.3074	0.1976	0.1287
$\boldsymbol{\beta_7}$	$\mathbf{-0.5868}$	$\mathbf{0.2041}$	$\mathbf{0.0068}$
$\boldsymbol{\beta_8}$	$\mathbf{-0.4721}$	$\mathbf{0.1780}$	$\mathbf{0.0120}$
$\boldsymbol{\beta_9}$	$\mathbf{-0.5865}$	$\mathbf{0.1839}$	$\mathbf{0.0030}$
$\boldsymbol{\beta_{10}}$	$\mathbf{-0.5826}$	$\mathbf{0.1843}$	$\mathbf{0.0032}$
β_{11}	-0.1544	0.2003	0.4459

4.5.3 Discussion

The findings show that the GOLLE regression model is appropriate for explaining the dengue fever weekly cases in the Federal District. Table 15 provides parameter estimates for the GOLLE regression model, which becomes:

$$\begin{aligned}\hat{\lambda}_i = \exp\Big(&-9.1118 + 0.5459\, m_{i1} + 1.0759\, m_{i2} + 1.6690\, m_{i3} + 1.7019\, m_{i4} \\ &+ 1.4481\, m_{i5} - 0.5868\, m_{i7} - 0.4721\, m_{i8} - 0.5865\, m_{i9} - 0.5826\, m_{i10}\Big).\end{aligned} \tag{4.10}$$

The following discussion examines the systematic structure using January as the month of reference.

Interpretations for systematic structure $\boldsymbol{\lambda}$

- Except for the covariates m_6 and m_{11}, referring to the months July and December, all other covariates are significant at a 5% level of significance;
- The months of February to June have positive estimates and it is significant. The distribution is positively skewed with heavy tails, and these months have a higher positive impact on the baseline distribution's shape parameter, possibly pulling the mean higher.

This may be seen in Figure 25(b), which shows a extreme event in that window data scenario. Several factors can contribute to extreme occurrences in dengue fever cases, including outbreaks, epidemics, severe cases, unusual patterns, and others. The event that occurred between May and June is considered atypical when compared to the period's historical average;

- August to November have negative values, which have a negative effect on the baseline distribution's shape parameter, shifting the average dengue cases lower. During this period, the Federal District experiences a drought that corroborates the findings[14]

4.6 CONCLUDING REMARKS

The Chapter defined the *generalized odd log-logistic exponential* distribution (QOSHJA; MUÇA, 2018) and (AFIFY et al., 2021) and introduced the regression model with a shape systematic structure to investigate dengue fever weekly cases in the Federal District in 2022. Some mathematical properties are presented, the parameters are estimated by the maximum likelihood method and the consistency criterion is evaluated by means of Monte Carlo simulations. The consistency of the MLEs of the regression model is evaluated by means of simulations. Some global influence measures and residual analysis are addressed to investigate the fit of the new model.

Some important discoveries are addressed. Except for the July and December months, the remaining months are significant. February to June exhibit positive estimations, suggesting a positive impact on the baseline distribution's shape parameter, possibly displacing the mean of cases higher. August to November experience a drought, supporting the negative estimates during this period, which have a negative effect on the shape parameter, altering the average dengue fever cases downward.

The epidemiology data set demonstrated that the novel model is more versatile than some nested and competing models. As a result, the suggested model improves understanding of dengue fever cases in the Federal District, as well as extreme events observed during the study period.

[14] <https://portal.inmet.gov.br/uploads/notastecnicas/Estado-do-clima-no-Brasil-em-2022-OFICIAL.pdf>

5 A NOVEL BIMODAL GAMMA REGRESSION MODEL WITH AGRICULTURAL APPLICATION

RESUMO

A distribuição gamma log-logística *odd* generalizada (CORDEIRO et al., 2017) é exibida e um novo modelo bimodal de regressão com duas componentes sistemáticas de forma e escala é introduzido para estudar dados agrícolas de batata yacon de uma pesquisa do Peru. Uma nova representação linear é demonstrada e algumas propriedades matemáticas são apresentadas. O método de máxima verossimilhança é usado para estimar os parâmetros e simulações são realizadas para investigar a precisão dos estimadores. Os resíduos *deviance* e a análise de diagnóstico são discutidos para avaliar o ajuste do modelo. A abordagem proposta identifica com eficiência os fatores essenciais que afetam a forma e escala da distribuição. Além disso, algumas descobertas valiosas são abordadas e a nova distribuição fornece um modelo bimodal alternativo adequado para estudos futuros em conjunto de dados não-Gaussianos.

Palavras-chaves: Dados agrícolas. Diagnósticos. Distribuição gama. Família log-logística *odd* generalizada. Máxima verossimilhança. Modelo bimodal. Modelo de regressão. Simulação.

ABSTRACT

The generalized odd log-logistic gamma distribution (CORDEIRO et al., 2017) is shown and a new bimodal regression model with two systematic components of shape and scale is introduced to study yacon potato agricultural data from a Peruvian survey. A new linear representation is demonstrated and some mathematical properties are presented. The maximum likelihood method is used to estimate the parameters and simulations are carried out to investigate the accuracy of the estimators. Deviation residuals and diagnostic analysis are discussed to assess the fit of the model. The proposed approach efficiently identifies the essential factors affecting the shape and scale of the distribution. In addition, some valuable findings are addressed and the new distribution provides an alternative bimodal model suitable for future studies on non-Gaussian data sets.

Keywords: Agricultural data. Bimodal model. Diagnostics. Gamma distribution. Generalized odd log-logistic family. Maximum likelihood. Regression model. Simulation.

5.1 INTRODUCTION

In statistics, the gamma distribution is a common continuous probability distribution used to model non-negative rvs. Numerous applications have been identified for this adaptable and frequently utilized distribution in a variety of fields, including finance (TRAN; KUKAL, 2022), engineering (KHAMEES et al., 2022) and the natural sciences (WANG; PAL, 2022).

While the gamma distribution is widely used, developing other statistical distributions has gained more attention recently. These innovative distributions often have more flexible and novel structures, providing researchers with important tools to more thoroughly investigate a range of data sets. Therefore, there will be a lot of research done in the future on the design of new statistical distributions.

Several recent studies have explored regression analysis across a range of contexts. For instance, a log-odd log-logistic Weibull regression model (**??**) is used to analyze heart transplant data, while Alizadeh et al. (2018) developed a heteroscedastic regression model to examine long-term survival in gastric adenocarcinoma data. In finance, Vasconcelos et al. (2019) introduced a new odd log-logistic generalized inverse Gaussian regression model. Additionally, Altun et al. (2021) proposed a new type II half logistic-G family with various properties, regression models, system reliability and applications. At last, Altun et al. (2022) explored the statistical properties, characterizations and regression modeling of the Gudermannian generated family.

Considering this scenario, the primary goal is to describe the *generalized odd log-logistic gamma* (GOLLΓ) distribution, capable of effectively modeling data with two distinct modes, or bimodal data. By fitting a model that accurately captures the underlying distribution of the data, more precise estimates for key quantities of interest can be obtained.

The second goal of this Chapter is to propose a regression model based on the GOLLΓ distribution that can effectively capture explanatory variables. In order to attain this goal, a novel analysis approaches are applied to agricultural data (yacon potatoes), with the MLE method used for inference. Such an approach is especially useful when the answer variable is influenced by many explanatory variables. Monte Carlo simulations are used to evaluate the accuracy of the proposed model.

To verify the validity of the results, the model assumptions are carefully examined and possible influential observations are investigated, this combined with residual analysis. Furthermore, envelope plots are used to illustrate the models' fit. This study intends to contribute to the existing literature by introducing a new regression model and shedding light on the link

between explanatory variables and the response variable.

This Chapter is well-structured and organized into several Sections. Section 5.2 provides a review of the GOLLΓ distribution, while Section 5.3 demontrates a new linear representation and some mathematical properties. MLEs are determined and a simulation study is conducted to verify their accuracy. The new GOLLΓ bimodal regression model is developed in Section 5.4 and the consistency of the estimators is examined. Section 5.5 demonstrates the superiority of the new regression model compared to other competitive models for explaining an agricultural dataset and presents some findings. The model is the best fit to explain the current data, according to diagnostic analysis and deviance residuals. Finally, Section 5.6 concludes with some remarks.

5.2 THE GOLLΓ DISTRIBUTION

The gamma distribution (shape parameter $p > 0$ and scale parameter $\mu > 0$) has the cdf and the pdf expressed, respectively (for $x > 0$), as:

$$G(x; p, \mu) = \gamma_1\left(p, \frac{x}{\mu}\right) \tag{5.1}$$

and

$$g(x; p, \mu) = \frac{1}{\mu^p \Gamma(p)} x^{p-1} \mathrm{e}^{-\frac{x}{\mu}}, \tag{5.2}$$

where $\gamma_1(p, x/\mu) = \frac{1}{\Gamma(p)} \int_0^{x/\mu} t^{p-1}\mathrm{e}^{-t} dt$ is the incomplete gamma function ratio and $\Gamma(\cdot)$ is the gamma function.

To obtain the GOLLΓ cdf, Equation (5.1) is substituted into Equation (2.2), as follows:

$$F(y; \alpha, \theta, p, \mu) = \frac{\gamma_1\left(p, \frac{y}{\mu}\right)^{\alpha\theta}}{\gamma_1\left(p, \frac{y}{\mu}\right)^{\alpha\theta} + \left[1 - \gamma_1\left(p, \frac{y}{\mu}\right)^{\theta}\right]^{\alpha}} \tag{5.3}$$

and the corresponding pdf is obtained by inserting Equation (5.2) into Equation (2.3) as:

$$f(y; \alpha, \theta, p, \mu) = \frac{\alpha\,\theta\, y^{p-1}\mathrm{e}^{-y/\mu}\gamma_1\left(p, \frac{y}{\mu}\right)^{\alpha\theta-1}\left[1 - \gamma_1\left(p, \frac{y}{\mu}\right)^{\theta}\right]^{\alpha-1}}{\mu^p\Gamma(p)\left\{\gamma_1\left(p, \frac{y}{\mu}\right)^{\alpha\theta} + \left[1 - \gamma_1\left(p, \frac{y}{\mu}\right)^{\theta}\right]^{\alpha}\right\}^2}. \tag{5.4}$$

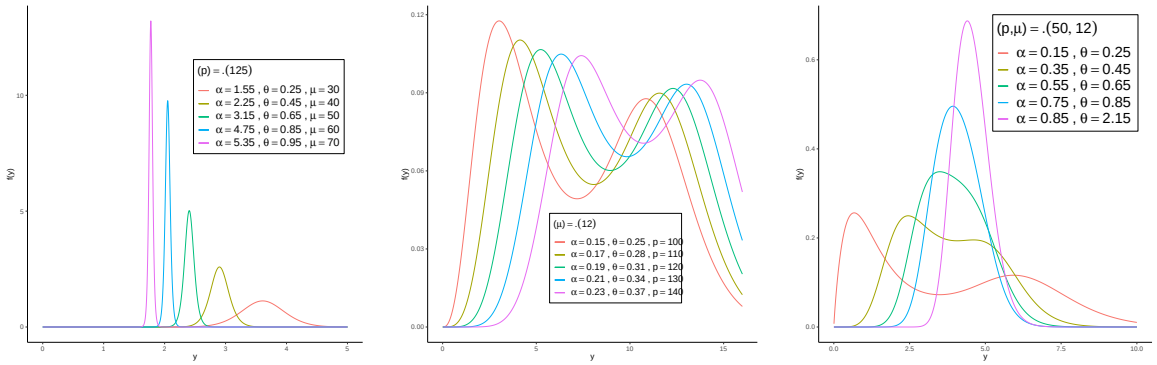
The GOLLΓ model includes six distinct distributions, which are summarized in Table 16. This indicates that the new class has the ability to fit data across multiple distribution forms.

The GOLLΓ distribution's pdf and hrf are shown in Figures 30 and 31, respectively. The flexibility of the GOLLΓ distribution to generate a variety of hazard rate shapes, such as the

Table 16 – Submodels associated to the GOLLΓ family of distributions.

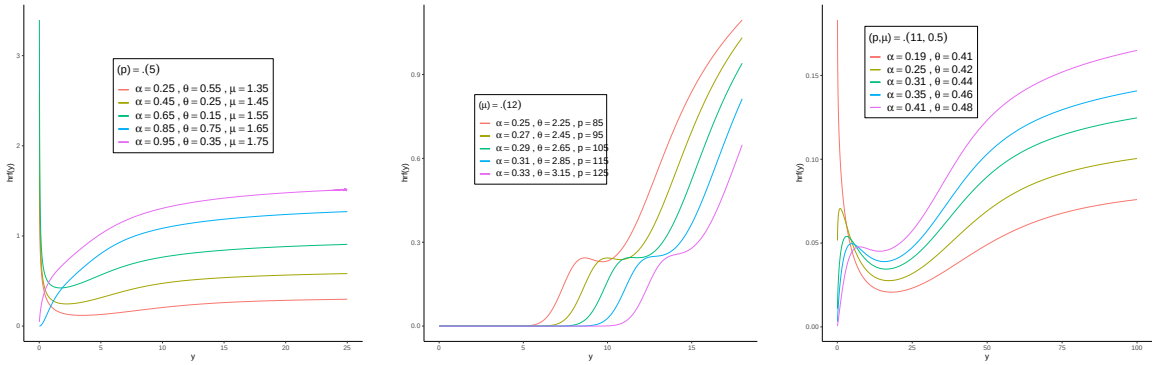
α	θ	p	μ	**Submodel**
-	-	1	-	Generalized odd log-logistic exponential (GOLLE) distribution
-	1	-	-	Odd log-logistic gamma (OLLΓ) distribution
1	1	-	-	Odd log-logistic exponential (OLLE) distribution
1	-	-	-	Exponentiated-gamma (Exp-Γ) distribution (NADARAJAH; GUPTA, 2007)
1	-	1	-	Exponentiated-exponential (Exp-E) distribution
1	1	-	-	Gamma (Γ) distribution

bathtub, inverse J-shape, increasing-decreasing, and other forms, is one of its key characteristics. It is superior to the hrf of the gamma, which presents simple shapes (constant, increasing, decreasing, etc). As a result, the model is useful for modeling data sets with a wide range of hazard rate patterns. Additionally, this model provides bimodality and skewness forms for a given parameter vector, as illustrated in Figure 30.



(a) (b) (c)

Figure 30 – GOLLΓ pdf for selected values.



(a) (b) (c)

Figure 31 – GOLLΓ hrf for selected values..

5.3 MAIN PROPERTIES

Since there is no mathematical closed-form for the GOLLΓ distribution, a novel linear combination of generalized gamma (GG) densities is used to represent the density function.

With shape parameters $p, \beta > 0$ and a scale parameter $\mu > 0$, the cdf of the GG distribution (STACY, 1962) is represented by:

$$\Lambda(x; p, \mu, \beta) = \gamma_1\left(p, \left(\frac{x}{\mu}\right)^{\beta}\right) \tag{5.5}$$

and its corresponding density function defined, as follows:

$$\lambda(x; p, \mu, \beta) = \frac{\beta}{\mu^{p\beta}\Gamma(p)} x^{p\beta-1} \exp\left[-\left(\frac{x}{\mu}\right)^{\beta}\right] \tag{5.6}$$

and clearly, the gamma pdf can be derived from (5.6) by setting $\beta = 1$.

Theorem 3. *The GOLLΓ density (5.4) can be represented linearly using the GG density:*

$$f(y; \alpha, \theta, p, \mu) = \sum_{k,m=0}^{\infty} z_{k,m}\, \lambda(y; p^*, \mu, 1), \tag{5.7}$$

with parameters $p^* = p^*(p, k, m) = k + p(m+1)$, μ, $\beta = 1$ and $z_{k,m}$ defined by the quantities below:

$$z_{k,m} = \frac{(m+1)\mu^{k+p(m+1)}\Gamma(k + p(m+1))}{\mu^p\Gamma(p)}\, b_k\, s_{k,m}.$$

Proof. The GOLLΓ density (5.4) can be expressed using the linear representation in Equation (2.5), as:

$$f(y; \alpha, \theta, p, \mu) = \sum_{k=0}^{\infty} b_k h_{k+1}(y; p, \mu), \tag{5.8}$$

where $h_{k+1}(y; p, \mu)$ is the Exp-Γ density with power parameter $(k+1)$ presented, as:

$$h_{k+1}(y; \mu, \lambda) = (k+1)\left[\frac{1}{\mu^p\Gamma(p)} x^{p-1} \mathrm{e}^{-\frac{x}{\mu}}\right]\left[\frac{1}{\mu^p\Gamma(p)} x^{p-1} \mathrm{e}^{-\frac{x}{\mu}}\right]^k$$

Moreover, the following expression was obtained by applying the power series expansion for the incomplete gamma function ratio:

$$\gamma_1(p, y/\mu) = \sum_{m=0}^{\infty} c_m\, y^{m+p}, \tag{5.9}$$

where $c_m = c_m(\alpha, \theta) = \frac{(-1)^m}{(m+p)\mu^{m+p}\Gamma(p)m!}$ (for $m \geq 0$).

Using the Equation for power series raised to powers found in Section 0.314 (GRADSHTEYN; JEFFREY; RYZHIK, 1996), the following expression holds:

$$\gamma_1(p, y/\mu)^k = \sum_{m=0}^{\infty} s_{k,m}\, y^{m+kp}, \tag{5.10}$$

where $s_{k,0} = c_0^k$ and $s_{k,m} = (mc_0)^{-1} \sum_{r=1}^{k} [(k+1)r - m] c_r\, s_{k,m-r}$ (for $m = 1, 2, \ldots$).

Finally, the GOLLΓ density can be expressed as a linear combination of GG densities:

$$\begin{aligned} f(y) &= \sum_{k,m=0}^{\infty} \frac{(k+1)}{\mu^p \Gamma(p)} b_k s_{k,m}\, y^{m+p(k+1)-1}\, \mathrm{e}^{-\frac{y}{\mu}} \\ &= \sum_{k,m=0}^{\infty} z_{k,m}\, \lambda(y; p^*, \mu, 1), \end{aligned} \tag{5.11}$$

with two shared parameters, μ, $\beta = 1$ and the third parameter p^*, where is determined by $p^* = p^*(p, k, m) = m + p(k+1)$ and the coefficients $z_{k,m}$, which depend on the preceding quantities, are expressed by:

$$z_{k,m} = \frac{(k+1)\mu^{m+p(k+1)}\Gamma(m + p(k+1))}{\mu^p \Gamma(p)}\, b_k\, s_{k,m}.$$

□

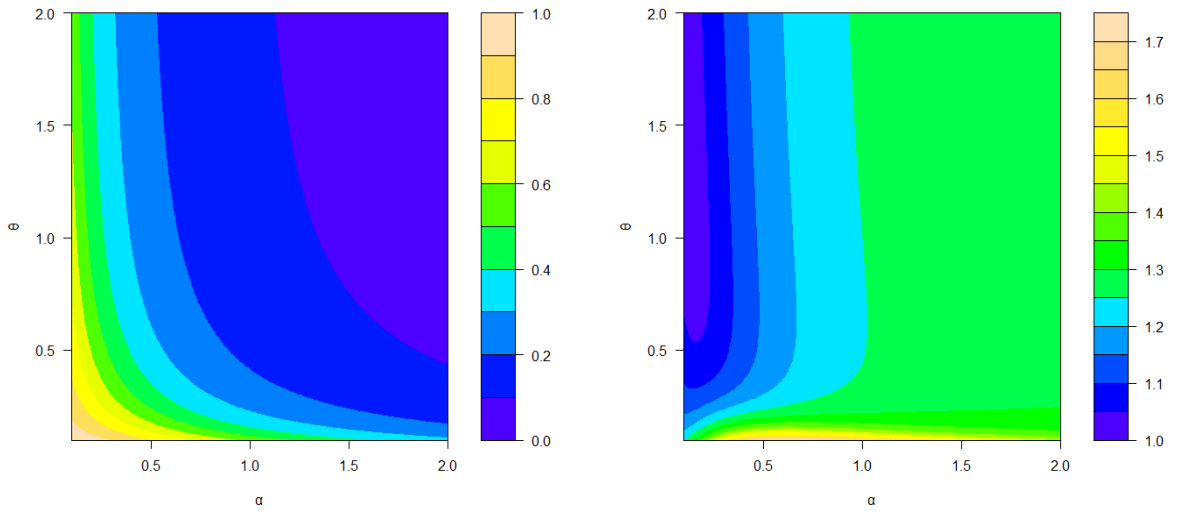
The linear representation (5.7) is the main result of this Section and a valuable tool in deriving mathematical properties for the GOLLΓ distribution utilizing well-established GG properties, see (STACY; MIHRAM, 1965) and (LAWLESS, 1980).

5.3.1 Quantile function

The qf of a rv Y can be derived using the gamma distribution with a parent distribution as described in Equation (2.6), as follows:

$$Q(u) = \mu \gamma_1^{-1}(p, \varepsilon_{\alpha,\theta}(u)). \tag{5.12}$$

With $p = 3.15$ and $\mu = 0.45$, Figure 32 presents Galton's skewness and Moors' kurtosis for a set of values of α and θ. This clearly illustrates how skewness and kurtosis act as functions of parameters. When both α and θ increase, the skewness decreases. Kurtosis increases for a region when both α and θ parameters increase.



(a) **(b)**

Figure 32 – GOLLΓ distribution (a) Galton's skewness. (b) Moor's kurtosis.

5.3.2 Moments

Using the moments of the GG densities, the moments of the GOLLΓ is calculated.

Theorem 4. *The nth moment of the GOLLΓ distribution is defined by:*

$$\mu'_n = \mathbb{E}(Y^n) = \mu^n \sum_{k,m=0}^{\infty} \frac{\Gamma(m+n+p(k+1))}{\Gamma(m+p(k+1))} z_{k,m}. \tag{5.13}$$

Proof. The proof is straightforward applying the Theorem (3) and using the basic properties of GG distribution, see (STACY; MIHRAM, 1965) and (LAWLESS, 1980). □

5.3.3 Generation function

The gf of the GOLLΓ distribution is shown using the gf of GG densities.

Theorem 5. *The mgf of the GOLLΓ density can be expressed, as:*

$$M_Y(t) = \sum_{n=0}^{\infty} \sum_{k,m=0}^{\infty} \frac{(t\mu)^n}{n!} \frac{\Gamma(m+n+p(k+1))}{\Gamma(m+p(k+1))} z_{k,m}. \tag{5.14}$$

Proof. The proof is straightforward applying the Theorem (3) and using the basic properties of GG distribution, see (STACY; MIHRAM, 1965) and (LAWLESS, 1980). □

5.3.4 Estimation

The log-likelihood function for the parameter vector $\boldsymbol{\psi} = (\alpha, \theta, p, \mu)^\top$ can be found based on the observed values $y_1, \ldots, y_n$ from the GOLLΓ distribution, as:

$$\begin{aligned} l_n(\boldsymbol{\psi}) &= n\log(\alpha\theta) - np\log(\mu) - n\log[\Gamma(p)] - (p-1)\sum_{i=1}^{n}\log(y_i) \\ &- \frac{1}{\mu}\sum_{i=1}^{n} y_i + (\alpha\theta - 1)\sum_{i=1}^{n}\log\left[\gamma_1\left(p, \frac{y_i}{\mu}\right)\right] + (\alpha-1)\sum_{i=1}^{n}\log\left[1-\gamma_1\left(p, \frac{y_i}{\mu}\right)^{\theta}\right] \\ &+ (\alpha-1)\sum_{i=1}^{n}\log\left[1-\gamma_1\left(p, \frac{y_i}{\mu}\right)^{\theta}\right] \\ &- 2\sum_{i=1}^{n}\log\left\{\gamma_1\left(p, \frac{y_i}{\mu}\right)^{\alpha\theta} + \left[1-\gamma_1\left(p, \frac{y_i}{\mu}\right)^{\theta}\right]^{\alpha}\right\}. \end{aligned} \tag{5.15}$$

For simplicity, let be

$$A_i(p,\mu) = A_i = \gamma_1\left(p, \frac{y_i}{\mu}\right).$$

The components of the score vector can be expressed, as follows:

$$\begin{aligned} U_\alpha &= \frac{n}{\alpha} + \theta\sum_{i=1}^{n}\log(A_i) + \sum_{i=1}^{n}\log\left(1-{A_i}^{\theta}\right) \\ &- 2\sum_{i=1}^{n}\frac{\theta\log(A_i){A_i}^{\alpha\theta} + (1-{A_i}^{\theta})^{\alpha}\log(1-{A_i}^{\theta})}{{A_i}^{\alpha\theta} + (1-{A_i}^{\theta})^{\alpha}}, \end{aligned}$$

$$\begin{aligned} U_\theta &= \frac{n}{\theta} + \alpha\sum_{i=1}^{n}\log(A_i) - (\alpha-1)\sum_{i=1}^{n}\frac{{A_i}^{\theta}\log(A_i)}{1-{A_i}^{\theta}} \\ &+ \sum_{i=1}^{n}\frac{\alpha{A_i}^{\alpha\theta}\log(A_i) + (1-{A_i}^{\alpha})^{\theta}\log(1-{A_i}^{\alpha})}{{A_i}^{\alpha\theta} + (1-{A_i}^{\alpha})^{\theta}}, \end{aligned}$$

$$\begin{aligned} U_p &= -n\log(\mu) - n\psi(k) - \sum_{i=1}^{n}\log(y_i) - (\alpha\theta-1)\sum_{i=1}^{n}\frac{\partial A_i/\partial p}{A_i} \\ &- \theta(\alpha-1)\sum_{i=1}^{n}\frac{A_i^{\theta-1}\partial A_i/\partial p}{1-A_i^{\theta}} - 2\alpha\theta\sum_{i=1}^{n}\frac{[A_i^{\alpha\theta-1} - (1-A_i^{\theta})^{\alpha-1}A_i^{\theta-1}]\partial A_i/\partial p}{A_i^{\alpha\theta} + (1-A_i^{\theta})^{\alpha}} \end{aligned}$$

and

$$\begin{aligned} U_\mu &= \frac{np}{\mu} + \frac{1}{\mu^2}\sum_{i=1}^{n} y_i + (\alpha\theta-1)\sum_{i=1}^{n}\frac{\partial A_i/\partial\mu}{A_i} - \theta(\alpha-1)\sum_{i=1}^{n}\frac{A_i^{\theta-1}\partial A_i/\partial\mu}{1-A_i^{\theta}} \\ &- 2\alpha\theta\sum_{i=1}^{n}\frac{[A_i^{\alpha\theta-1} - (1-A_i^{\theta})^{\alpha-1}A_i^{\theta-1}]\partial A_i/\partial\mu}{A_i^{\alpha\theta} + (1-A_i^{\theta})^{\alpha}}. \end{aligned}$$

The formulas for the quantities $\partial\gamma_1(\cdot)/\partial p$ and $\partial\gamma_1(\cdot)/\partial\mu$ are provide in Marciano et al. (2012), as:

$$\frac{\partial\gamma_1(p, y/\mu)}{\partial p} = \Gamma(p)\psi(\gamma) - \log(y/\mu)\gamma(p, y/\mu) - G_{2,3}^{3,0}\left(\begin{matrix}1,\ 1\\ 0,\ 0,\ p\end{matrix}\middle|\ y/\mu\right),$$

where $G_{2,3}^{3,0}\left(.\ \middle|\ \cdot\right)$ is a particular case of the Meijer G-function[15] and

$$\frac{\partial\gamma_1(p, y/\mu)}{\partial\mu} = \frac{y^p\exp(-y/\mu)}{\mu^{p-1}\Gamma(p)}.$$

Using the *optim* routine, Equation (5.15) is numerically maximized to calculate the MLE of the vector $\hat{\boldsymbol{\psi}}$.

5.3.5 Simulation study

One-thousand samples of sizes $n = \{50, 100, 250\}$ are generated for each of the two sets of true parameters $(0.25, 2.00, 0.50, 0.35)$ for simulation 1 and $(0.38, 0.75, 1.50, 2.35)$ for simulation 2 in order to evaluate the estimators' accuracy. Table 17 reports the AEs, biases, and MSEs, for $\boldsymbol{\epsilon} = (\alpha, \theta, p, \mu)$. The consistency criteria are held since the AEs converge to the true values and the biases and MSEs tend to zero as the sample size increases.

Table 17 – Simulations results for GOLLΓ distribution

Simulation 1									
Par	**n = 50**			**n = 100**			**n = 250**		
	AE	**Bias**	**MSE**	**AE**	**Bias**	**MSE**	**AE**	**Bias**	**MSE**
α	0.289	0.039	0.031	0.282	0.032	0.016	0.274	0.024	0.007
θ	2.206	0.206	4.098	2.118	0.118	2.286	2.172	0.172	1.579
p	0.611	0.111	0.810	0.446	-0.054	0.339	0.329	-0.171	0.160
μ	1.027	0.677	0.944	0.836	0.486	0.451	0.743	0.393	0.239
Simulation 2									
Par	**n = 50**			**n = 100**			**n = 250**		
	AE	**Bias**	**MSE**	**AE**	**Bias**	**MSE**	**AE**	**Bias**	**MSE**
α	0.441	0.061	0.137	0.401	0.021	0.045	0.391	0.011	0.020
θ	2.273	1.523	15.210	2.100	1.350	10.241	1.694	0.944	4.526
p	5.137	3.627	84.410	3.130	1.630	17.763	2.116	0.616	4.792
μ	3.690	1.340	11.268	2.713	0.363	2.905	2.144	-0.206	0.790

[15] <https://functions.wolfram.com/HypergeometricFunctions/MeijerG/>

5.4 THE GOLLΓ REGRESSION MODEL.

The GOLLΓ regression model's systematic component establishes a relationship between the response variable and one or more covariates. It is typical practice in statistical modeling to take into account the variation in parameter values across several observations. Thus, for $i = 1, \ldots, n$, the regression model is specified by two systematic components for p_i and μ_i, as follows:

$$p_i = \exp(\boldsymbol{x}_{i1}^{\top}\boldsymbol{\beta_1}) \quad \text{and} \quad \mu_i = \exp(\boldsymbol{x}_{i2}^{\top}\boldsymbol{\beta_2}). \tag{5.16}$$

Here, $\boldsymbol{x_{ij}} = (x_{ij1}, \cdots, x_{ijc_j})$ represents the observations on c_j known regressors ($j = 1, 2$) and $\boldsymbol{\beta} = (\beta_{j0}, \cdots, \beta_{jc})^{\top}$ are vectors of length $(c_j + 1)$ of unknown coefficients functionally independent. The log-linear link function is assumed to be twice continuously differentiable and plays a key role in specifying the relationship between the response variable and the covariates.

5.4.1 Estimation

With the exception of the vector parameters of the gamma distribution, the components of the score vector of U_α and U_θ are the same Equations given in Subsection (5.3.4). The score components of the vector parameters p_i and μ_i are defined for the purpose to add the regression part, respectively, as follows:

$$\begin{aligned} U_{p_i} &= \sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_1}} g(x_i;p_i)}{g(x_i;p_i)} + (\alpha\theta - 1)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_1}} G(x_i;p_i)}{G(x_i;p_i)} + \theta(1-\alpha)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_1}} G(x_i;p_i) G(x_i;p_i)^{\theta-1}}{1 - G(x_i;p_i)^{\theta}} \\ &- 2\sum_{i=1}^{\infty} \partial_{\boldsymbol{\beta_1}} G(x_i;p_i) \frac{G(x_i;p_i)^{\alpha\theta-1} - G(x_i;p_i)^{\theta-1}[1 - G(x_i;p_i)^{\theta}]^{\alpha-1}}{G(x_i;p_i)^{\alpha\theta} + [1 - G(x_i;p_i)^{\theta}]^{\alpha}} \end{aligned}$$

and

$$\begin{aligned} U_{\mu_i} &= \sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_2}} g(x_i;\mu_i)}{g(x_i;\mu_i)} + (\alpha\theta - 1)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_2}} G(x_i;\mu_i)}{G(x_i;\mu_i)} + \theta(1-\alpha)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_2}} G(x_i;\mu_i) G(x_i;\mu_i)^{\theta-1}}{1 - G(x_i;\mu_i)^{\theta}} \\ &- 2\sum_{i=1}^{\infty} \partial_{\boldsymbol{\beta_2}} G(x_i;\mu_i) \frac{G(x_i;\mu_i)^{\alpha\theta-1} - G(x_i;\mu_i)^{\theta-1}[1 - G(x_i;\mu_i)^{\theta}]^{\alpha-1}}{G(x_i;\mu_i)^{\alpha\theta} + [1 - G(x_i;\mu_i)^{\theta}]^{\alpha}}, \end{aligned}$$

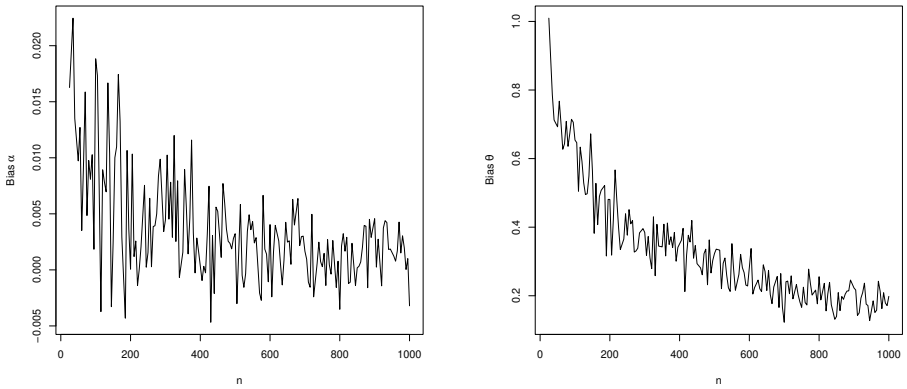
where $\partial_{\boldsymbol{\beta_1}} g(x_i;p_i) = \partial_{p_i} g(x_i;p_i)\partial_{\boldsymbol{\beta_1}} p_i(x_i;\boldsymbol{\beta_1})$ and $\partial_{\boldsymbol{\beta_1}} G(x_i;p_i) = \partial_{p_i} G(x_i;p_i)\partial_{\boldsymbol{\beta_1}} p_i(x_i;\boldsymbol{\beta_1})$ denotes the derivatives of the parameter p_i and $\partial_{\boldsymbol{\beta_2}} g(x_i;\mu_i) = \partial_{\mu_i} g(x_i;\mu_i)\partial_{\boldsymbol{\beta_2}} \mu_i(x_i;\boldsymbol{\beta_2})$ and $\partial_{\boldsymbol{\beta_2}} G(x_i;\mu_i) = \partial_{\mu_i} G(x_i;\mu_i)\partial_{\boldsymbol{\beta_2}} \mu_i(x_i;\boldsymbol{\beta_2})$ denotes the derivatives of the parameter μ_i using the chain rule.

The MLE $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ of the regression model is calculated setting the score equations $U_\alpha = U_\theta = U_{p_i} = U_{\mu_i} = 0$ using an iterative method algorithm to find roots or using the *optim* routine.

5.4.2 Simulation study

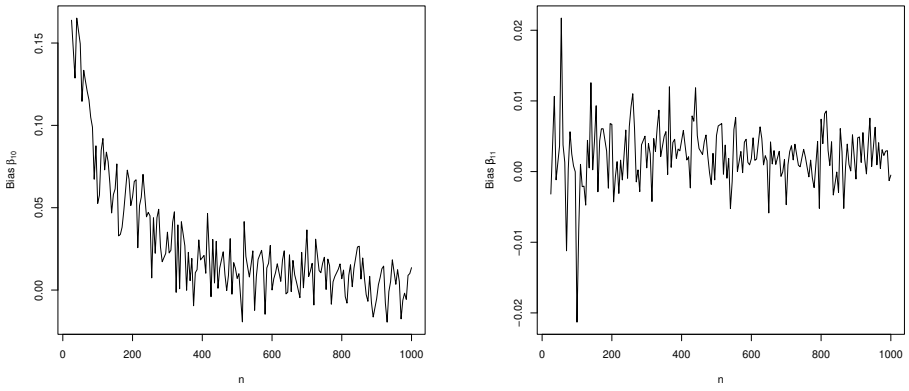
GOLLΓ qf was used to simulate 1,000 samples with varying sample sizes ($n = 25, \ldots, 1,000$) and fixed parameter values ($\alpha = 0.50, \theta = 3.00, \beta_{10} = 0.75, \beta_{11} = 0.40, \beta_{20} = 0.20, \beta_{21} = 3.00$). The MLEs for the parameters were estimated for each generated sample. Following that, the biases, MSEs, ALs and CPs, were determined for each parameter estimate. This study enabled the MLEs' accuracy and precision to be assessed across a range of sample sizes and parameter values.

The Figures 33-42 show how these measures behave as a function of sample size. As n increase, the biases, MSEs and ALs converge towards zero, showing that the MLEs are consistent. Furthermore, as n increases, the CPs approach 0.95, showing that the estimators are becoming more accurate. The biases of the parameters β_{11} and β_{21} present are near zero as the sample size grows in an oscillator behavior, nonetheless, this form does not affect the estimative of all other parameters.



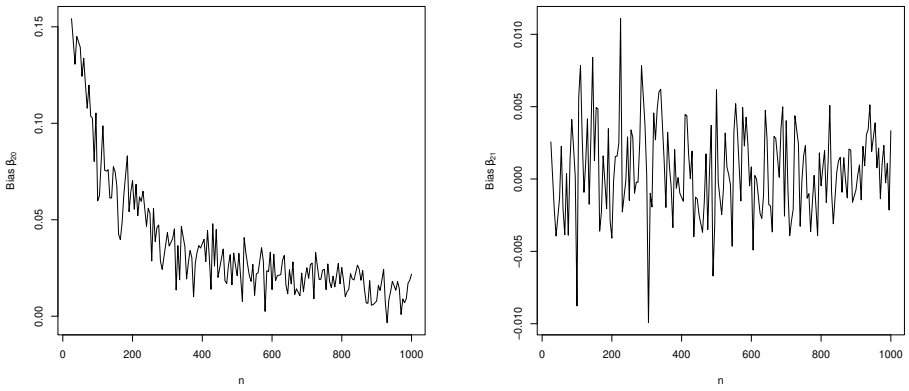
(a) (b)

Figure 33 – Biases versus sample size from GOLLΓ regression model.



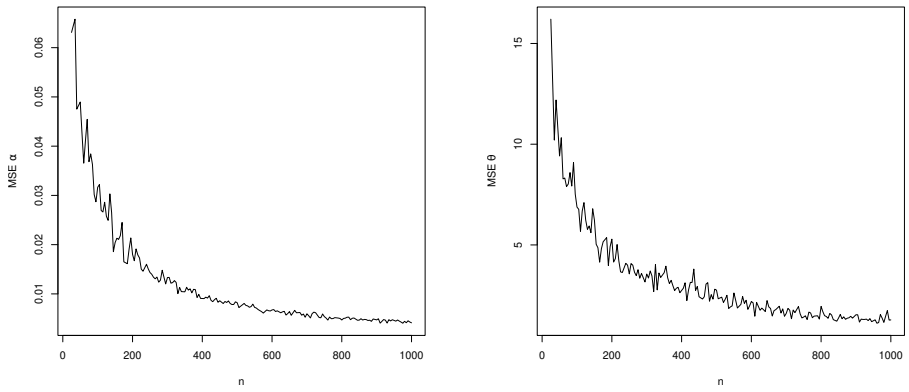
(a) (b)

Figure 34 – Biases versus sample size from GOLLΓ regression model.



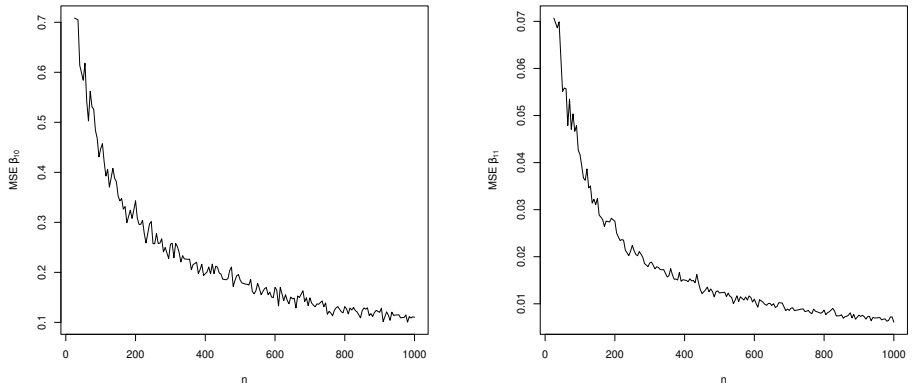
(a) (b)

Figure 35 – Biases versus sample size from GOLLΓ regression model.



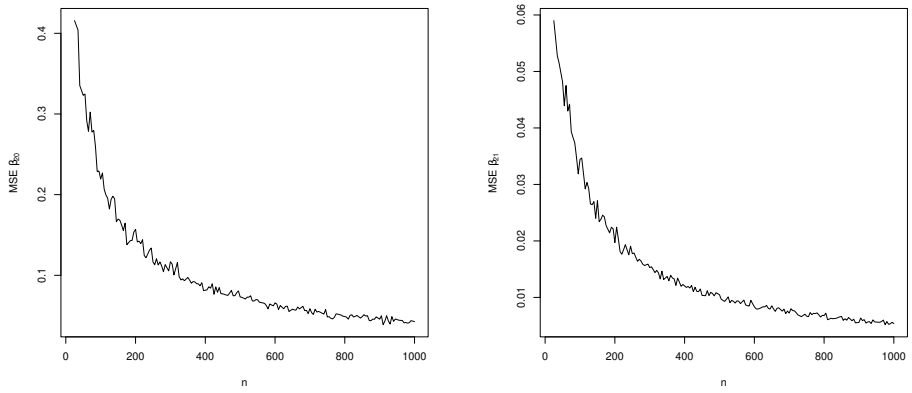
(a) (b)

Figure 36 – MSEs versus sample size from GOLLΓ regression model.



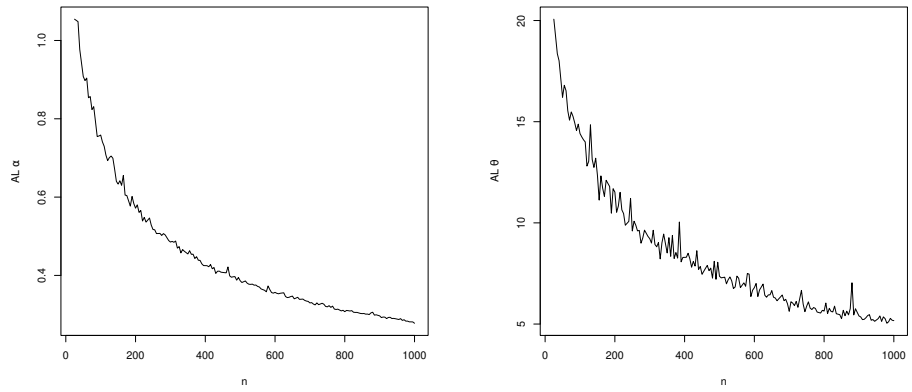
(a) (b)

Figure 37 – MSEs versus sample size from GOLLΓ regression model.



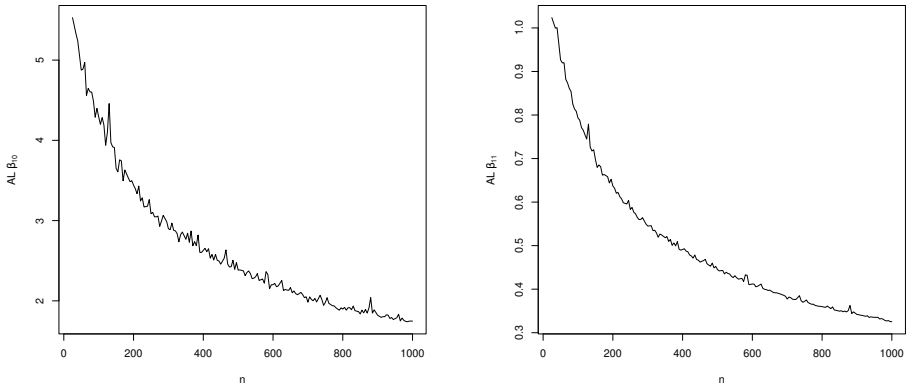
(a) (b)

Figure 38 – MSEs versus sample size from GOLLΓ regression model.



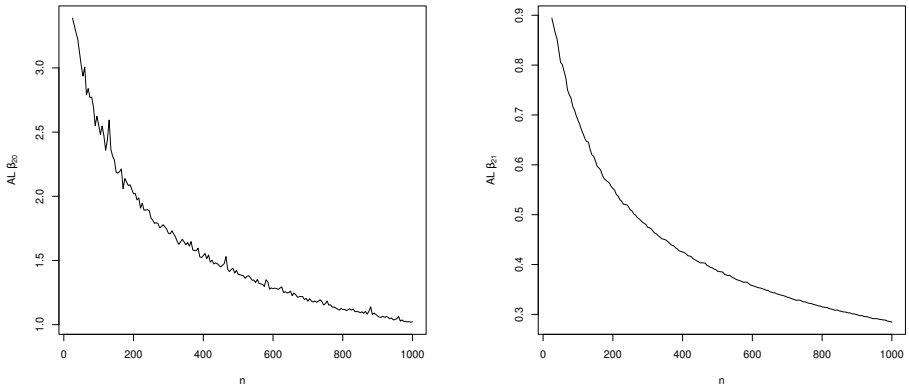
(a) (b)

Figure 39 – ALs versus sample size from GOLLΓ regression model.



(a) (b)

Figure 40 – ALs versus sample size from GOLLΓ regression model.



(a) (b)

Figure 41 – ALs versus sample size from GOLLL regression model.

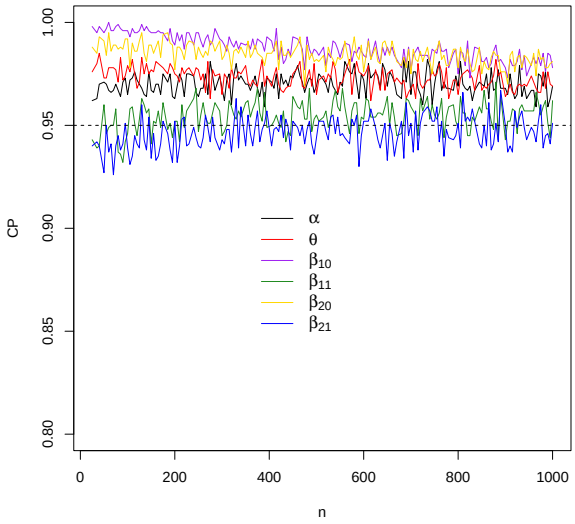


Figure 42 – CPs versus sample size from GOLLΓ regression model.

5.5 APPLICATION: AGRICULTURAL (YACON POTATO) DATA

The GOLLΓ model is compared with nested models and other alternative distributions as indicating in Table 18.

Table 18 – Competitive distributions compared to the GOLLΓ distribution.

Distribution	**Reference**
Kumaraswamy-gamma (KwΓ)	(CORDEIRO; CASTRO, 2011)
New Weibull-Gamma (NWΓ)	(KLAKATTAWI, 2019)
New flexible-gamma (NFΓ)	(TAHIR; HUSSAIN; CORDEIRO, 2022)

The densities are (for $x > 0$), respectivelly:

$$F_{\mathsf{Kw}\Gamma}(x) = 1 - \{1 - G(x)^{\alpha}\}^{\theta},$$

$$F_{\mathsf{NW}\Gamma}(x) = 1 - \exp\left\{-\alpha\left[\frac{G(x)}{1-G(x)}\right]^{\theta}\right\}$$

and

$$F_{\mathsf{NF}\Gamma}(x) = 1 - [1 - G(x)]^{G(x)},$$

where all the parameters are positive and $G(x)$ is defined in Equation (5.1). The *goodness.fit* function computes the MLEs (SEs in parenthesis) for all fitted models using the BFGS approach.

5.5.1 Dataset definition

The application examines yacon data provided from the *agricolae* package (Felipe de Mendiburu; Muhammad Yaseen, 2020) in R Core Team (2021) to illustrate the usefulness of the new regression model over other competitive models. Ivan Manrique and Carolina Tasso contributed the data from the International Potato Center's (IPC, Peru) experimental field in 2003.

The dataset has 432 observations of yacon potato (*Smallanthus sonchifolius*), a native plant of the Peruvian Andes. The effect of the covariates on the response variable degree brix (sugar concentration) was investigated. Figure 43 presents the bimodal behavior of the data.

To achieve this purpose, our focus was on the following variables:

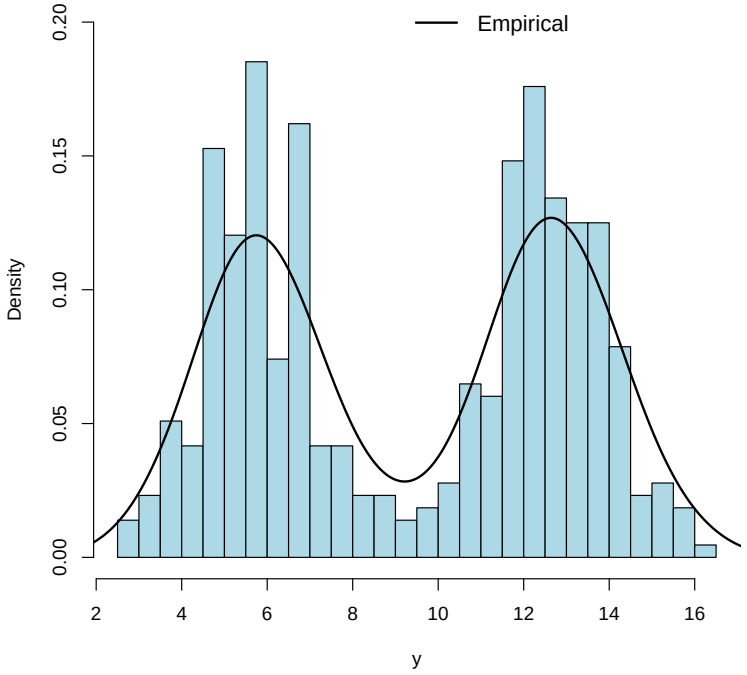


Figure 43 – Histogram and empirical density of yacon potato data.

- y_i: Degree brix (DBR) - measures the density or sugar concentration of solutions (response variable);
- x_{i1}: Sucrose (SUCR) - pergentage;
- x_{i2}: Dummy variable 1 (LOC1) - location - Cajamarca $= 1$, Lima $= 0$, Oxapampa $= 0$;
- x_{i3}: Dummy variable 2 (LOC2) - location - Cajamarca $= 0$, Lima $= 1$, Oxapampa $= 0$;
- x_{i4}: Fructo-oligosaccharides (FOS) - pergentage;
- x_{i5}: Fructose (FRUC) - pergentage.

Vasconcelos et al. (2021) explored the data set using the odd log-logistic exponential Gaussian (OLLExGa) regression model to explain the association between the response variable and location. The goal is to add to the existing literature on this topic by providing new and helpful insights into the numerous factors that influence the response variable.

The descriptive statistics from the dataset are presented in Table 19. The degree brix mean is 9.43, with higher standard deviation indicating greater variability in the data. The near zero skewness suggests a central distribution, while negative kurtosis indicates a distribution with

thinner tails and a flatter peak compared to a normal distribution. Based on these findings, it is expected that SUCR will relate positively with DBR, given both are measurements of sugar content. Similar behavior could be expected from the variables FOS and FRUC.

Table 19 – Descriptive statistics of yacon potato data.

Variable	**Statistics**						
	Mean	**Median**	**SD**	**Skewness**	**Kurtosis**	**Min.**	**Max.**
DBR	9.43	10.45	3.67	-0.05	-1.54	2.90	16.10
SUCR	18.07	18.67	8.14	-0.24	-0.43	-1.36	37.30
FOS	45.43	48.40	20.19	-0.64	-0.33	0.70	79.20
FRUC	17.82	14.98	13.28	0.56	-0.68	1.50	54.15

5.5.2 Results

Table 20 displays the findings of the fitted models and the GoF measures, jointly with the MLEs and SEs in parenthesis. The GOLLΓ distribution was found to have the best fit to the data, except for the KS measure (OLLΓ - 0.096). The proposed distribution is also the best fit model when compared to the odd log-logistic exponential Gaussian (OLLExGa) model introduced by Vasconcelos et al. (2021) (AIC: 2142.956 versus 2149.156). The estimated pdfs and cdfs presented in Figure 44 for the fitted models reinforce this.

Table 20 – Findings from the fitted models of yacon potato data.

Model	α	θ	p	μ	**AIC**	**BIC**	W*	A*	**KS**
GOLLΓ	**0.097**	**0.790**	**160.132**	**17.890**	**2142.956**	**2159.230**	**1.008**	**5.378**	**0.128**
	(0.004)	**(0.024)**	**(0.055)**	**(0.003)**					
OLLΓ	0.084	1	166.068	18.400	2156.692	2168.898	1.207	6.636	0.096
	(0.003)	(-)	(0.022)	(0.008)					
EΓ	1	0.014	298.962	17.700	2301.724	2313.930	2.848	15.224	0.174
	(-)	(0.005)	(96.082)	(5.578)					
Γ	1	1	5.906	0.626	2350.633	2358.770	3.266	17.847	0.178
	(-)	(-)	(0.391)	(0.043)					
GOLLE	0.522	52.497	1	0.537	2353.793	2365.998	2.806	15.790	0.205
	(0.065)	(20.802)	(-)	(0.059)					
OLLE	2.742	1	1	0.077	2390.677	2398.814	3.601	19.901	0.165
	(0.107)	(-)	(-)	(0.002)					
EE	1	8.286	1	0.290	2367.808	2375.945	3.323	18.289	0.184
	(-)	(0.809)	(-)	(0.012)					
KwΓ	0.213	64.742	14.973	0.186	2337.605	2353.879	3.173	17.239	0.163
	(0.202)	(60.570)	(14.171)	(0.182)					
NWΓ	2.907	15.784	0.996	1.484	2332.191	2348.465	3.118	16.917	0.159
	(0.357)	(10.124)	(0.653)	(0.872)					
NFΓ	-	-	2.821	0.357	2371.753	2379.890	3.414	18.791	0.178
	(-)	(-)	(0.196)	(0.025)					

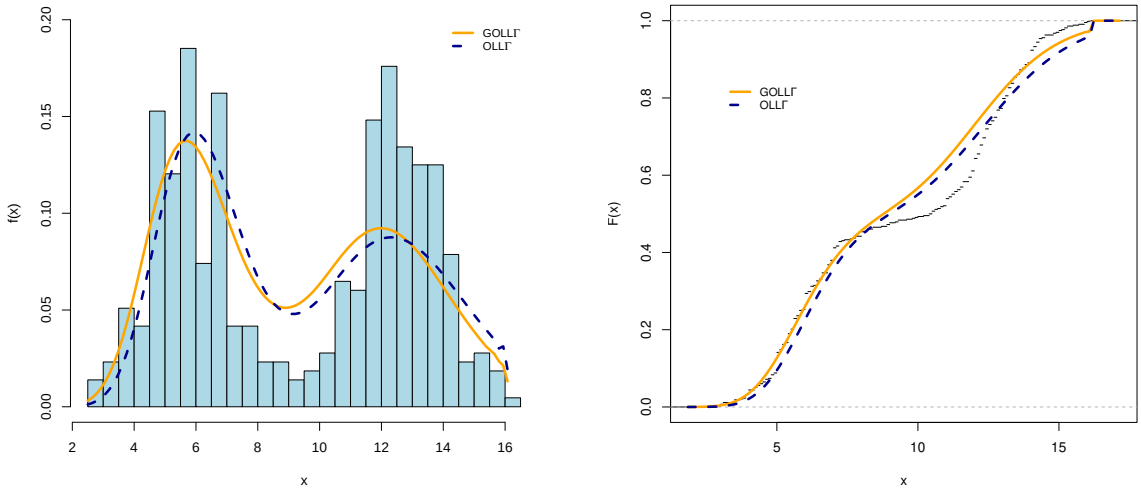


Figure 44 – Fitted models of yacon potato data. (a) Histogram and estimated pdfs. (b) Empirical and estimated cdfs.

To compare the novel distribution with its nested models, a LR test was used. The results in Table 21 show that including more parameters is significant for modeling the existing data.

Table 21 – LR tests of yacon potato data.

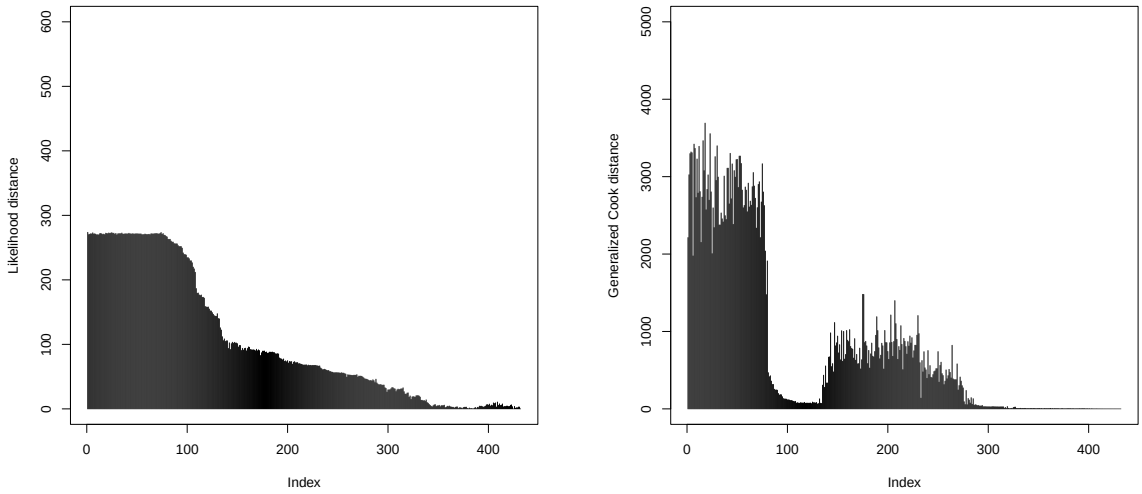
Models	**Statistic w**	**p-value**
GOLLΓ vs Γ	211.677	< 0.0001
GOLLΓ vs EΓ	160.768	< 0.0001
GOLLΓ vs OLLΓ	15.736	< 0.0001
GOLLΓ vs EE	228.852	< 0.0001
GOLLΓ vs OLLE	251.721	< 0.0001
GOLLΓ vs GOLLE	212.837	< 0.0001

Next, one focused on the systematic components of the Equation (5.16) (for $i = 1, \ldots, 432$), as:

$$\begin{aligned} \hat{p}_i &= \exp(\beta_{10} + \beta_{11}\, x_{i1} + \beta_{12}\, x_{i2} + \beta_{13}\, x_{i3} + \beta_{14}\, x_{i4} + \beta_{15}\, x_{i5}), \\ \hat{\mu}_i &= \exp(\beta_{20} + \beta_{21}\, x_{i1} + \beta_{22}\, x_{i2} + \beta_{23}\, x_{i3} + \beta_{24}\, x_{i4} + \beta_{25}\, x_{i5}). \end{aligned} \tag{5.17}$$

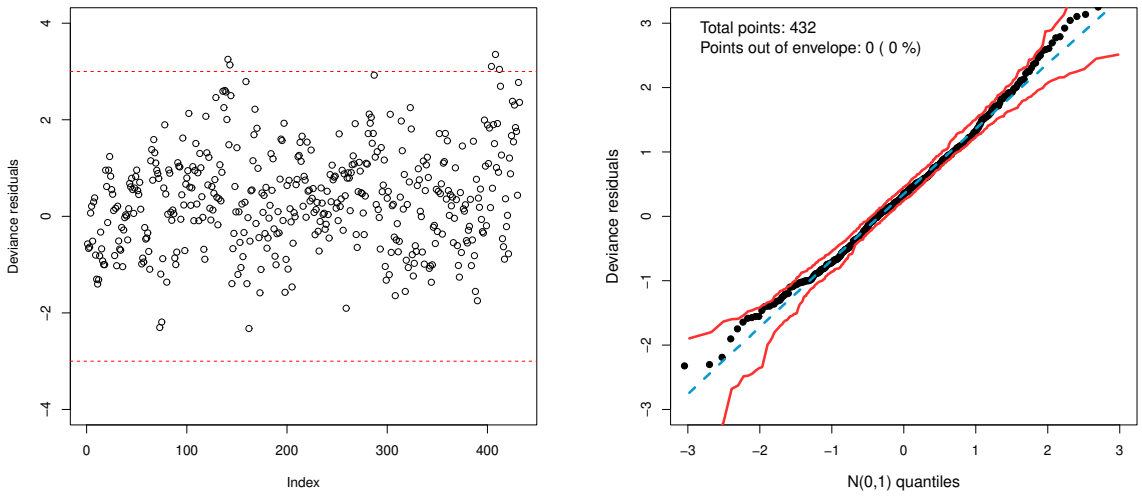
Examining Figure 45, some observations are found to impact the model, proving the usefulness of LD and GCD measures in identifying potentially influential observations. To further assess the suitability of the new regression model, Figure 46(a) shows the index plot of the deviance residuals, indicating random behavior, besides some observations. Additionally, Figure 46(b) presents the normal probability plot with simulated envelope, suggesting that the

GOLLΓ regression model is suitable for analyzing these data. These results imply that the possible influential observations had no effect on the model.



(a) **(b)**

Figure 45 – The GOLLΓ regression model. (a) LD. (b) GCD.



(a) **(b)**

Figure 46 – The GOLLΓ regression model. (a) Deviance residual index. (b) Simulated envelope.

Table 22 displays the results with the MLEs, SEs and p-values obtained from fitting the new regression model proposed.

Table 22 – Fitted GOLLΓ regression of yacon potato data.

	MLEs	**SEs**	p**-values**		**MLEs**	**SEs**	p**-values**
$\boldsymbol{\beta_{10}}$	**-3.1373**	**0.9194**	**0.0007**	$\boldsymbol{\beta_{20}}$	**-2.0724**	**0.5350**	**0.0001**
β_{11}	-1.3455	1.4679	0.3599	$\boldsymbol{\beta_{21}}$	**-1.9263**	**0.7287**	**0.0085**
$\boldsymbol{\beta_{12}}$	**-3.8563**	**0.5227**	**<0.0001**	$\boldsymbol{\beta_{22}}$	**-0.9061**	**0.1759**	**<0.0001**
$\boldsymbol{\beta_{13}}$	**-2.9413**	**0.2212**	**<0.0001**	$\boldsymbol{\beta_{23}}$	**-1.0075**	**0.1564**	**<0.0001**
$\boldsymbol{\beta_{14}}$	**5.6532**	**0.9318**	**<0.0001**	$\boldsymbol{\beta_{24}}$	**2.6710**	**0.5870**	**<0.0001**
$\boldsymbol{\beta_{15}}$	**10.2580**	**1.3435**	**<0.0001**	$\boldsymbol{\beta_{25}}$	**5.6557**	**0.8984**	**<0.0001**

5.5.3 Discussion

The model checks show that the agricultural (yacon potato) data can be adequately explained by the GOLLΓ regression model. By utilizing the parameter estimates presented in Table 22, the GOLLΓ regression model is transformed into:

$$\begin{aligned}\hat{p}_i &= \exp(-3.1373 - 3.8563\,x_{i2} - 2.9413\,x_{i3} + 5.6532\,x_{i4} + 10.2580\,x_{i5}),\\ \hat{\mu}_i &= \exp(-2.0724 - 1.9263\,x_{i1} - 0.9061\,x_{i2} - 1.0075\,x_{i3} + 2.6710\,x_{i4} + 0.8984\,x_{i5}).\end{aligned} \tag{5.18}$$

Based on the parameter estimates provided in Table 22, several interpretations can be derived:

- All parameter estimates, except for β_{11}, exhibit a significant level of significance at the 5% level;
- **Interpretation for systematic structure p**
  - The distribution is symmetrical under the mean (near zero skewness) with a light tail (negative kurtosis), which means that the negative coefficients of β_{12} and β_{13}, respectively, LOC1 and LOC2, suggest that as the covariates decrease, the shape of the baseline distribution decreases, shifting the average of degree brix.
  - The variables FOS and FRUC have significant and positive coefficients (β_{14} and β_{15}), indicating an effect on increasing the shape parameter and, subsequently, the skewness. As a possible result, the impact on the degree brix of the variables is a mean right displaced.
- **Interpretation for systematic structure μ**

- The negative coefficients of β_{21}, β_{22}, and β_{23}, respectively, for SUCR, LOC1, and LOC2, indicate a reduction of the scale parameter, which can indicate a decrease in kurtosis with a negative impact on the spread of the degree brix in relation to these variables;
- In the opposite effect, the β_{24} and β_{25} coefficients are positive, resulting in an increase in the scale parameter and, as a result, an increase in the kurtosis. This has potentially a positive impact on the spread of the degree brix.

5.6 CONCLUDING REMARKS

In this Chapter, the *generalized odd log-logistic gamma* distribution is described and a novel bimodal regression model with a shape and scale systematic structure is used to explore the agricultural yacon potato data from a study in Peru. A novel linear representation was demonstrated and some mathematical properties were presented. The parameters were estimated using the maximum likelihood method and demonstrated the consistency of the estimators through Monte Carlo simulations. Likewise, in the presented distribution, the key features of the regression model are displayed with simulations. To assess the goodness-of-fit measures, diagnostic checks are performed and deviance residuals are examined, which indicate the new model's suitability.

Some valuable findings are shown by performing the model. Except for the sucrose coefficient in the shape structure, the remaining are significant at the 5% level. The descriptive Table showed that the distribution is symmetrical under the mean (near zero skewness) with a light tail (negative kurtosis). For the shape systematic component, locations 1 and 2 have negative estimates that reduce the baseline's shape parameter, which suggests shifting the average degree brix lower. The proportions of fructo-oligosaccharides and fructose have positive estimates, indicating that they increase the shape parameter and, consequently, the skewness. As a consequence, pulling the mean of the sugar concentration higher. For the scale systematic component, sucrose, locations 1 and 2 have negative estimates, indicating a reduction of the scale parameter, and in the kurtosis, indicating a negative impact on the spread of the degree brix. In contrast, fructo-oligosaccharides and fructose increase the scale parameter, resulting in an increase in kurtosis. Therefore, there is a likely positive impact on the spread of the sugar concentration.

The proposed model showed more flexibility than some competitive models and prior investigation models. The new bimodal model provides better insights to explain factors that affect the shape and scale parameters and serves as an alternative model to other research.

6 A WEIBULL REGRESSION MODEL IN PRESENCE OF BIMODALITY WITH APPLICATION TO WIND ENERGY GENERATION

RESUMO

A distribuição Weibull log-logística *odd* generalizada (CORDEIRO et al., 2017) é descrita e um novo modelo bimodal de regressão com dois componentes estruturais é indicado para averiguar dados do total de geração de energia eólica no Brasil. É apresentado algumas características matemáticas, os parâmetros determinados usando a máxima verossimilhança e as simulações de Monte Carlo confirmaram a suposição de consistência. Por meio de simulações, usando algumas medidas, os MLEs do modelo de regressão são avaliados. Medidas de diagnóstico de influência global e análise de resíduos foram usadas para identificar observações influentes e discrepantes. Para a janela temporal estudada, a regressão introduzida identificou os meses com impacto nos parâmetros de forma e escala. Os principais resultados indicaram que o modelo de regressão proposto é eficaz e são abordadas algumas interpretações para melhor compreender o conjunto de dados.

Palavras-chaves: Bimodalidade. Distribuição Weibull. Energia eólica. Família log-logística *odd* generalizada. Máxima verossimilhança. Modelo de Regressão. Simulação.

ABSTRACT

The generalized odd log-logistic Weibull distribution (CORDEIRO et al., 2017) is described and a new bimodal regression model with two structural components is indicated to investigate data on total wind power generation in Brazil. Some mathematical characteristics are presented, the parameters were determined using maximum likelihood and Monte Carlo simulations confirmed the consistency assumption. Through simulations, using some measures, the MLEs of the regression model are evaluated. Diagnostic measures of global influence and residual analysis were used to identify influential and discrepant observations. For the time window studied, the regression introduced identified the months with an impact on the shape and scale parameters. The main results indicated that the proposed regression model is effective and some interpretations are addressed to better understand the dataset.

Keywords: Bimodality. Generalized odd log-logistic family. Maximum likelihood. Regression model. Simulation. Weibull distribution. Wind energy.

6.1 INTRODUCTION

Renewable energy sources are gaining progress in the worldwide effort to combat climate change and reduce dependency on fossil fuels. Unlike harmful fuels, these resources utilize nature's power to supply clean energy while self-regenerating over time. Biomass, derived from organic materials such as plants and waste; geothermal energy, which uses the Earth's interior heat; hydroelectric power, generated by utilizing the force of flowing water, both in rivers and waves, providing a continuous and adaptable energy, are some examples for green energy. Nevertheless, wind energy generation stands out as a significant contributor to the move toward cleaner and more sustainable electricity generation.

In light of this, Delina (2022), for example, investigates the feasibility of harvesting Hong Kong's offshore wind energy, strategic implementation measures and associated environmental challenges. Roga et al. (2022) examines wind energy technology, explores possible futures and suggests solutions to increase the efficiency of wind generation systems. Moreover, the environmental impact, economic concerns and energy implications of wind energy generation are discussed (MSIGWA; IGHALO; YAP, 2022). Finally, the climate uncertainties regarding the future of wind power across Europe's climate zones are examined (YANG; JAVANROODI; NIK, 2022).

The Weibull distribution has been widely applied in a wide variety of fields. Ortega, Cruz e Cordeiro (2019) utilized the log-odd logistic-Weibull regression model under informative censoring to analyze the survival time of patients with chronic leukemia. Moreover, Ishaq e Abiodun (2020) introduced the Maxwell-Weibull distribution applied in an economic data set of exchange rates of Nigerian Naira to Japanese Yen. In addition, Vila e Çankaya (2022) defined a bimodal Weibull distribution and modeled six real data sets, which outperformed compared to competitive models. Further, Sayibu, Luguterah e Nasiru (2024) presented a new six-parameter Weibull distribution that contains several sub-models and demonstrated its applicability in lifetime data.

Based on this, the Chapter studies the total daily wind energy generation in Brazilian territory. The country offers wide areas suitable for wind farm installation, with the Northeast standing out for its good climatic conditions with continuous and strong winds, making it a significant center for wind energy generation. Wind energy's future outlook is critical in the Brazilian energy grid, with the potential for development in other regions of Brazil and continuous investment in more efficient and sustainable wind technology.

Therefore, the study offers a novel regression model based on the *generalized odd log-logistic Weibull* (GOLLW) distribution (CORDEIRO et al., 2017). This distribution is extremely adaptable when it comes to representing bimodal data. The parameters are determined using maximum likelihood and the accuracy of the GOLLW regression is assessed using Monte Carlo simulations. Some local influence measures are used to verify potential outlier observations and deviance residuals are used to verify some assumptions.

The rest of the Chapter is organized as follows. Section 6.2 presents the GOLLW distribution features. Some mathematical properties are addressed and the maximum likelihood method is used to estimate the parameters. The analysis of the accuracy of estimates is done by means of Monte Carlo simulations in Section 6.3. The new bimodal regression model with two systematic components was constructed in Section 6.4 and the MLEs of the parameters were described. Simulations are used to assess the accuracy of the regression MLEs. Section 6.5 illustrates the new regression's usefulness using Brazil wind energy generation data. Some conclusions are provided in Section 6.6.

6.2 THE GOLLW DISTRIBUTION

For more than fifty years, the Weibull distribution has been widely employed in a variety of fields (physics, medicine, economics, sociology, etc). The distribution is defined by its cdf and pdf (for $x > 0$), as follows:

$$G(x;\mu,\lambda) = 1 - \mathrm{e}^{-\left(\frac{x}{\lambda}\right)^{\mu}} \tag{6.1}$$

and

$$g(x;\mu,\lambda) = \frac{\mu}{\lambda}\left(\frac{x}{\lambda}\right)^{\mu-1}\mathrm{e}^{-\left(\frac{x}{\lambda}\right)^{\mu}}, \tag{6.2}$$

respectivelly, where $\mu > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter.

Inserting Equation (6.1) into (2.2) follows that the GOLLW cdf is

$$F(y;\alpha,\theta,\mu,\lambda) = \frac{\left(1-\mathrm{e}^{-\left(\frac{x}{\lambda}\right)^{\mu}}\right)^{\alpha\theta}}{\left(1-\mathrm{e}^{-\left(\frac{x}{\lambda}\right)^{\mu}}\right)^{\alpha\theta} + [1-\left(1-\mathrm{e}^{-\left(\frac{x}{\lambda}\right)^{\mu}}\right)^{\theta}]^{\alpha}}. \tag{6.3}$$

The corresponding four-parameter pdf, plugging Equations (6.1) and (6.2) in (2.3), is defined, as:

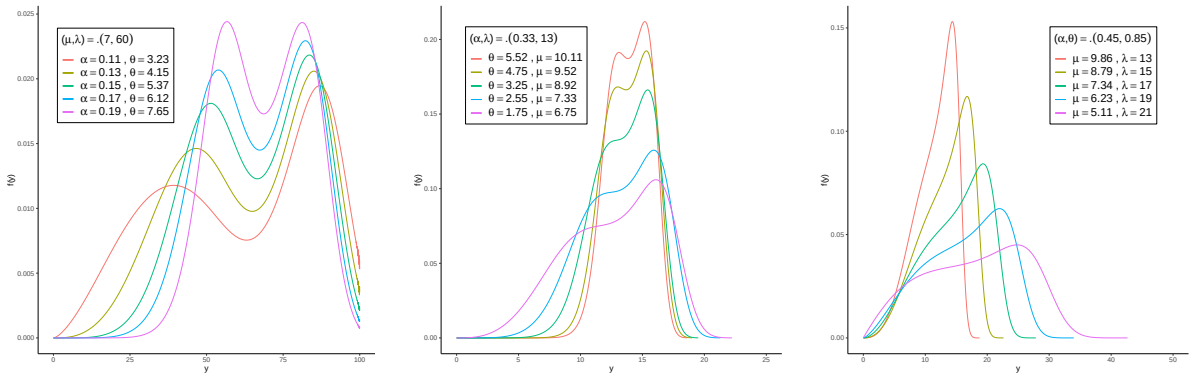
$$f(y;\alpha,\theta,\mu,\lambda) = \frac{\alpha\theta\,\frac{\mu}{\lambda}\,(\frac{x}{\lambda})^{\mu-1}\,\mathrm{e}^{-(\frac{x}{\lambda})^{\mu}}\,\left(1-\mathrm{e}^{-(\frac{x}{\lambda})^{\mu}}\right)^{\alpha\theta-1}\,[1-\left(1-\mathrm{e}^{-(\frac{x}{\lambda})^{\mu}}\right)^{\theta}]^{\alpha-1}}{\left\{\left(1-\mathrm{e}^{-log(\frac{x}{\lambda})^{\mu}}\right)^{\alpha\theta}+\left[1-\left(1-\mathrm{e}^{-(\frac{x}{\lambda})^{\mu}}\right)^{\theta}\right]^{\alpha}\right\}^{2}}. \tag{6.4}$$

The GOLLW contains 10 submodels, which are listed below. The new class proves its adaptability by fitting data to a variety of distribution forms and attributes.

Table 23 – Submodels associated to the GOLLW family of distributions.

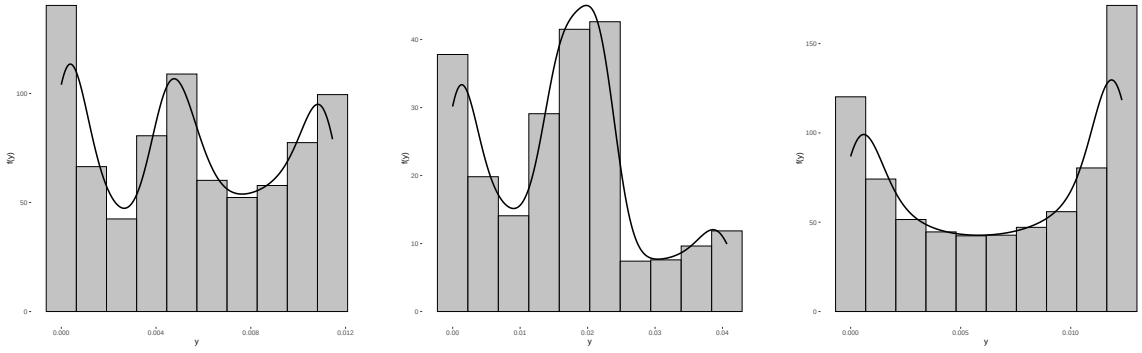
α	θ	μ	λ	**Submodel**
-	1	-	-	Odd log-logistic Weibull (OLLW) distribution (COORAY, 2006)
1	-	-	-	Exponentiated-Weibull (Exp-W) distribution (MUDHOLKAR; SRIVASTAVA, 1993a)
1	1	-	-	Weibull (W) distribution
-	-	1	-	Generalized odd log-logistic (GOLLE) dsitribution
-	1	1	-	Odd log-logistic exponential (OLLE) distribution
-	1	2	-	Odd log-logistic Rayleigh (OLLR) distribution
1	-	1	-	Exponentiated-exponential (Exp-E) dsitribution
1	-	2	-	Exponentiated-Rayleigh (Exp-R) distribution (SURLES; PADGETT, 2001)
1	1	1	-	Exponential (E) distribution
1	1	2	-	Raileigh (R) distribution

The GOLLW distribution's extra parameters allow great flexibility and a wide range of alternatives for density plots and histograms. Figure 47 shows many properties such as bi-modality and left/right skewness, which have more potential to fit a wide range of data than the Weibull distribution. Figure 48 illustrates multimodality and bathtub shape, these qualities demonstrate the new distribution's complexity and usefulness.



(a) (b) (c)

Figure 47 – GOLLW pdf for selected values.



(a) (b) (c)

Figure 48 – GOLLW histogram. (a) GOLLW(0.05,9.73,8,80). (b) GOLLW(0.10,4.13,9,40). (c) GOLLW(0.67,1.25,4,90).

6.3 MAIN PROPERTIES

A closed mathematical version of the GOLLW distribution is not possible, taking into consideration, the density of Y can be described as a linear representation using the Equation (2.5), as:

$$f(y;\alpha,\theta,\mu,\lambda) = \sum_{k=0}^{\infty} b_k h_{k+1}(y;\mu,\lambda),$$

where $h_{k+1}(y;\mu,\lambda)$ is the Exp-W density with power parameter $k+1$ presented, as:

$$h_{k+1}(y;\mu,\lambda) = (k+1)\left[\frac{\mu}{\lambda}\left(\frac{x}{\lambda}\right)^{\mu-1} \mathrm{e}^{-\left(\frac{x}{\lambda}\right)^{\mu}}\right]\left[1-\mathrm{e}^{-\left(\frac{x}{\lambda}\right)^{\mu}}\right]^{k}.$$

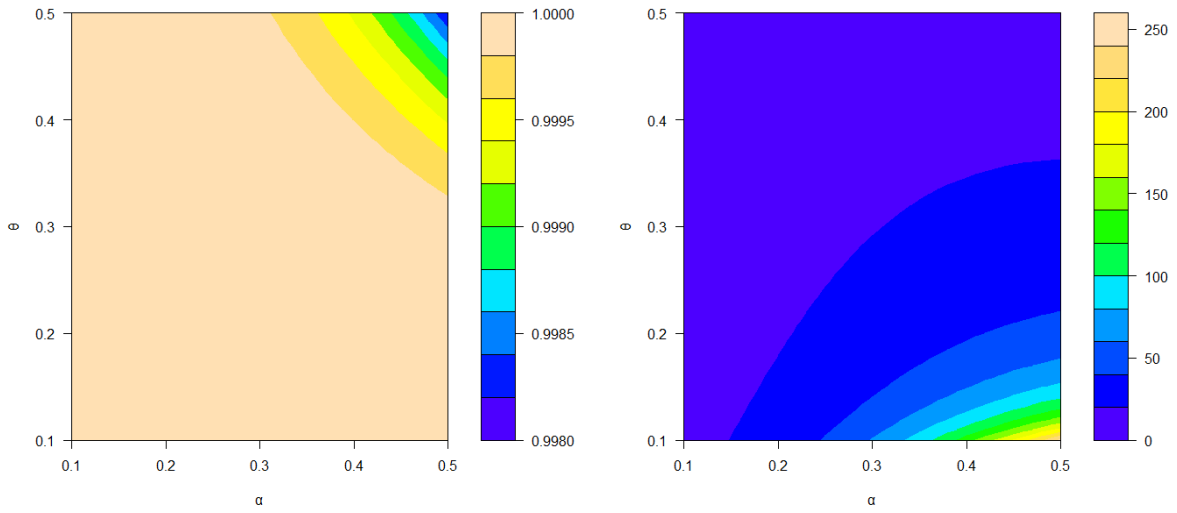
Hence, the GOLLW properties are obtained in a simple form by utilizing Exp-W properties, see (MUDHOLKAR; SRIVASTAVA, 1993b).

6.3.1 Quantile function

The qf of Y can be easily calculated using the Equation (2.6) and the qf of the Weibull distribution, as:

$$Q(u) = \lambda\left[\log\left(\frac{1}{1-\epsilon_{\alpha,\theta}(u)}\right)\right]^{1/\mu}. \tag{6.5}$$

For varying α and θ of the GOLLW(α,θ,0.25,0.85) distribution, the skewness and kurtosis measures are given in Figure 49, which show a decrease behavior for both measures. The skewness stabilizes when both parameters increase and the kurtosis drops to a zero region.



(a) (b)

Figure 49 – GOLLW distribution (a) Galton's skewness. (b) Moor's kurtosis.

6.3.2 Moments

Equation (2.7) calculates the moments of the GOLLW distribution based on the Exp-W distribution.

Theorem 6. *The nth ordinary moment of Y, for $n = 1, 2, 3, \ldots$, is given by:*

$$\mu'_n = \mathbb{E}(Y^n) = \sum_{k=0}^{\infty} a\lambda^n \Gamma\left(\frac{n}{\mu} + 1\right) \left\{1 + \sum_{i=1}^{\infty} c_i \left[(i+1)^{-n/\mu+1}\right]\right\} b_k, \tag{6.6}$$

where (for $i = 1, 2, 3, \ldots$)

$$c_i = (-1)^i \frac{(a-1)[(a-1)-1]\cdots[(a-1)-\overline{i-1}]}{i!}.$$

Proof. The proof is straightforward by applying Equation (6.3) and using the Exp-W moments in Choudhury (2005). □

6.3.3 Generation function

Equation (2.8) determines the mgf of the GOLLW distribution established on the gf of the Exp-W distribution.

Theorem 7. *The mgf of Y is*

$$M_Y(t) = \sum_{n,k,h} \frac{(-1)^h \Gamma(k+2)\,(t\lambda)^n}{j!\Gamma(k+1-h)\,(h+1)^{\frac{n}{\mu}+1}} \Gamma\left(\frac{n}{\mu} + 1\right) b_k. \tag{6.7}$$

Proof. The proof is straightforward by applying Eq. (6.3) and using the Exp-W mgf in Afify et al. (2018). □

6.3.4 Estimation

The MLEs of the parameters in $\boldsymbol{\psi} = (\alpha, \theta, \mu, \lambda)^\top$ are determined for a complete random sample $y_1, \ldots, y_n$ from the GOLLW distribution by maximizing:

$$\begin{aligned} l_n(\boldsymbol{\psi}) &= n\log(\alpha\theta) + n\log\left(\frac{\mu}{\lambda}\right) + (\mu-1)\sum_{i=1}^{n}\log\left(\frac{y_i}{\lambda}\right) \\ &- \sum_{i=1}^{n}\left(\frac{y_i}{\lambda}\right)^{\mu} + (\alpha\theta-1)\sum_{i=1}^{n}\log\left(1-\mathrm{e}^{-\left(\frac{y_i}{\lambda}\right)^{\mu}}\right) + (\alpha-1)\sum_{i=1}^{n}\log\left[1-\left(1-\mathrm{e}^{-\left(\frac{y_i}{\lambda}\right)^{\mu}}\right)^{\theta}\right] \\ &- 2\sum_{i=1}^{n}\log\left\{\left(1-\mathrm{e}^{-\left(\frac{y_i}{\lambda}\right)^{\mu}}\right)^{\alpha\theta} + \left[1-\left(1-\mathrm{e}^{-\left(\frac{y_i}{\lambda}\right)^{\mu}}\right)^{\theta}\right]^{\alpha}\right\}. \end{aligned} \tag{6.8}$$

Let be

$$A_i = A_i(\mu, \lambda) = 1 - \mathrm{e}^{-\left(\frac{y_i}{\lambda}\right)^{\mu}}$$

.

Thus, the score vector are

$$U_\alpha = \frac{n}{\alpha} + \theta\sum_{i=1}^{n}\log A_i + \sum_{i=1}^{n}\log(1-A_i) - 2\sum_{i=1}^{n}\frac{\theta A_i^{\alpha\theta}\log A_i + (1-A_i^{\theta})\log(1-A_i^{\theta})^{\alpha}}{A_i^{\alpha\theta} + (1-A_i^{\theta})^{\alpha}},$$

$$\begin{aligned} U_\theta &= \frac{n}{\theta} + \alpha\sum_{i=1}^{n}\log A_i + (1-\alpha)\sum_{i=1}^{n}\frac{A_i^{\theta}\log A_i}{1-A_i^{\theta}} \\ &- 2\alpha\sum_{i=1}^{n}\frac{A_i^{\alpha\theta}\log A_i - A_i^{\theta}(1-A_i^{\theta})^{\alpha-1}\log A_i}{A_i^{\alpha\theta} + (1-A_i^{\theta})^{\alpha}}, \end{aligned}$$

$$\begin{aligned} U_\mu &= \frac{n}{\mu} + \sum_{i=1}^{n}\log\left(\frac{y_i}{\lambda}\right) - \sum_{i=1}^{n}\log\left(\frac{y_i}{\lambda}\right)\left(\frac{y_i}{\lambda}\right)^{\mu} + (\alpha\theta-1)\sum_{i=1}^{n}\frac{\partial_\mu A_i}{A_i} \\ &+ \theta(1-\alpha)\sum_{i=1}^{n}\frac{\partial_\mu A_i \log A_i^{\theta-1}}{1-A_i^{\theta}} \\ &- 2\alpha\theta\sum_{i=1}^{n}\partial_\mu A_i\frac{A_i^{\alpha\theta-1} - A_i^{\theta-1}(1-A_i^{\theta})^{\alpha-1}}{A_i^{\alpha\theta} + (1-A_i^{\theta})^{\alpha}} \end{aligned}$$

and

$$\begin{aligned} U_\lambda &= -(\mu-2)\frac{n}{\lambda} + \frac{\mu}{\lambda}\sum_{i=1}^{n}\left(\frac{y_i}{\lambda}\right)^{\mu} + (\alpha\theta-1)\sum_{i=1}^{n}\frac{\partial_\lambda A_i}{A_i} + \theta(1-\alpha)\sum_{i=1}^{n}\frac{\partial_\lambda A_i \log A_i^{\theta-1}}{1-A_i^{\theta}} \\ &- 2\alpha\theta\sum_{i=1}^{n}\partial_\lambda A_i\frac{A_i^{\alpha\theta-1} - A_i^{\theta-1}(1-A_i^{\theta})^{\alpha-1}}{A_i^{\alpha\theta} + (1-A_i^{\theta})^{\alpha}}, \end{aligned}$$

where the quantities $\partial_\mu A_i$ and $\partial_\lambda A_i$ are, respectively:

$$\partial_\mu A_i = \log\left(\frac{y_i}{\lambda}\right)\left(\frac{y_i}{\lambda}\right)^\mu (1 - A_i) \quad \text{and} \quad \partial_\lambda A_i = -\frac{\mu}{\lambda}\left(\frac{y_i}{\lambda}\right)^\mu (1 - A_i).$$

The MLEs can be determined using any type of Newton-Raphson method by fixing the score vector equations $U_\alpha = U_\theta = U_\mu = U_\lambda = 0$ or using the *optim* procedure.

6.3.5 Simulation study

Monte Carlo simulations with $1,000$ samples of varied sizes are used to assess the accuracy of MLEs under two scenarios. The AEs, ABs and RMSEs are calculated for each sample size $(n = 50, 100, 200, 400, 800, 1,000)$, for $\boldsymbol{\epsilon} = (\alpha, \theta, \mu, \lambda)$. In both schemes, all parameter estimates are overestimated, however, when the sample size increases, the estimates are near the true value and the ABs and the MSEs are close to zero.

Table 24 – Simulations results for GOLLW distribution

scheme 1 - GOLLW(0.45,1.25,1.09,1.14)									
Par	n=50			n =100			n = 200		
	AE	AB	RMSE	AE	AB	RMSE	AE	AB	RMSE
α	0.536	0.089	0.416	0.525	0.075	0.323	0.509	0.059	0.265
θ	2.044	0.794	2.116	1.593	0.343	1.259	1.398	0.148	0.797
μ	1,548	0.458	1.312	1.310	0.220	0.649	1.217	0.127	0.405
λ	2.227	1.087	2.836	1.867	0.727	2.114	1.657	0.517	1.697
Par	n = 400			n = 800			n = 1000		
α	0.461	0.011	0.145	0.460	0.010	0.100	0.458	0.0084	0.0828
θ	1.345	0.095	0.512	1.287	0.037	0.363	1.274	0.0236	0.3068
μ	1.130	0.040	0.162	1.105	0.015	0.087	1.100	0.0090	0.0717
λ	1.281	0.141	0.732	1.210	0.070	0.396	1.191	0.0510	0.3078
scheme 2 - GOLLW(0.65,0.90,0.88,0.76)									
Par	n = 50			n = 100			n = 200		
	AE	AB	RMSE	AE	AB	RMSE	AE	AB	RMSE
α	0.441	0.111	0.374	0.410	0.080	0.299	0.382	0.052	0.227
θ	1.947	0.497	1.731	1.683	0.233	1.141	1.560	0.110	0.785
μ	2.991	1.741	5.797	1.856	0.606	2.658	1.468	0.218	1.091
λ	2.980	1.500	3.826	2.442	0.962	2.868	2.050	0.570	2.147
Par	n = 400			n = 800			n = 1000		
α	0.335	0.005	0.101	0.337	0.007	0.068	0.334	0.004	0.058
θ	1.545	0.095	0.515	1.479	0.029	0.370	1.482	0.032	0.331
μ	1.293	0.043	0.235	1.264	0.014	0.091	1.260	0.010	0.073
λ	1.585	0.106	0.713	1.542	0.062	0.387	1.519	0.039	0.291

6.4 THE GOLLW REGRESSION MODEL

Modeling both the shape and scale parameters of the distribution (6.4) of Y yields the novel heteroscedastic regression model. To introduce the structure, the parameters μ_i and λ_i are assumed to vary among observations, as shown below:

$$\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta_1}) \quad \text{and} \quad \lambda_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta_2}). \tag{6.9}$$

Here, $\boldsymbol{x_{ij}} = (x_{ij1}, \cdots, x_{ijc_j})$ represents the observations on c_j known regressors $(j = 1, 2)$, and $\boldsymbol{\beta} = (\beta_{j0}, \cdots, \beta_{jc})^\top$ are vectors of length $(c_j + 1)$ of functionally independent unknown coefficients. The log-linear link function is considered to be twice continuously differentiable and plays an important role in describing the relationship between the response variable and the covariates.

6.4.1 Estimation

Apart for the vector parameters of the Weibull distribution, the components of the score vector of U_α and U_θ are the identical Equations presented in Subsection (6.3.4). The score components of the vector parameters μ_i and λ_i are specified to add the regression part, as follows:

$$\begin{aligned} U_{\mu_i} &= \sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_1}} g(x_i;\mu_i)}{g(x_i;\mu_i)} + (\alpha\theta - 1)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_1}} G(x_i;\mu_i)}{G(x_i;\mu_i)} + \theta(1-\alpha)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_1}} G(x_i;\mu_i) G(x_i;\mu_i)^{\theta-1}}{1 - G(x_i;\mu_i)^{\theta}} \\ &- 2\sum_{i=1}^{\infty} \partial_{\boldsymbol{\beta_1}} G(x_i;\mu_i) \frac{G(x_i;\mu_i)^{\alpha\theta-1} - G(x_i;\mu_i)^{\theta-1}[1 - G(x_i;\mu_i)^{\theta}]^{\alpha-1}}{G(x_i;\mu_i)^{\alpha\theta} + [1 - G(x_i;\mu_i)^{\theta}]^{\alpha}} \end{aligned}$$

and

$$\begin{aligned} U_{\lambda_i} &= \sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_2}} g(x_i;\lambda_i)}{g(x_i;\lambda_i)} + (\alpha\theta - 1)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_2}} G(x_i;\lambda_i)}{G(x_i;\lambda_i)} + \theta(1-\alpha)\sum_{i=1}^{\infty} \frac{\partial_{\boldsymbol{\beta_2}} G(x_i;\lambda_i) G(x_i;\lambda_i)^{\theta-1}}{1 - G(x_i;\lambda_i)^{\theta}} \\ &- 2\sum_{i=1}^{\infty} \partial_{\boldsymbol{\beta_2}} G(x_i;\lambda_i) \frac{G(x_i;\lambda_i)^{\alpha\theta-1} - G(x_i;\lambda_i)^{\theta-1}[1 - G(x_i;\lambda_i)^{\theta}]^{\alpha-1}}{G(x_i;\lambda_i)^{\alpha\theta} + [1 - G(x_i;\lambda_i)^{\theta}]^{\alpha}}, \end{aligned}$$

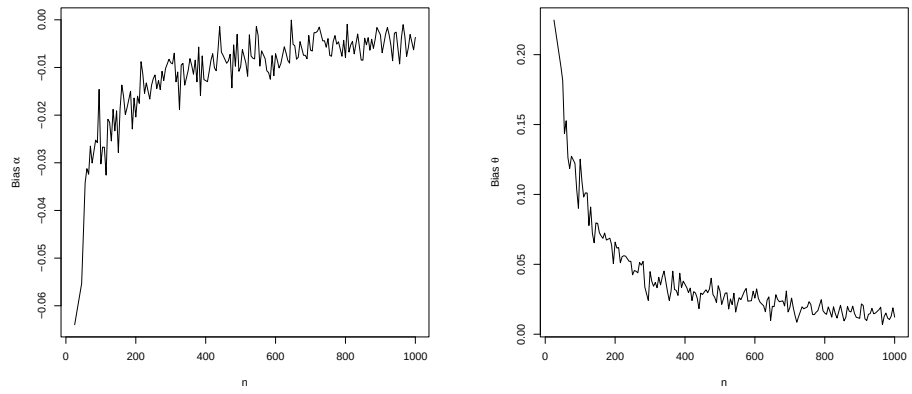
where $\partial_{\boldsymbol{\beta_1}} g(x_i;\mu_i) = \partial_{\mu_i} g(x_i;\mu_i)\partial_{\boldsymbol{\beta_1}}\mu_i(x_i;\boldsymbol{\beta_1})$ and $\partial_{\boldsymbol{\beta_1}} G(x_i;\mu_i) = \partial_{\mu_i} G(x_i;\mu_i)\partial_{\boldsymbol{\beta_1}}\mu_i(x_i;\boldsymbol{\beta_1})$ denotes the derivatives of the parameter μ_i and $\partial_{\boldsymbol{\beta_2}} g(x_i;\lambda_i) = \partial_{\lambda_i} g(x_i;\lambda_i)\partial_{\boldsymbol{\beta_2}}\lambda_i(x_i;\boldsymbol{\beta_2})$ and $\partial_{\boldsymbol{\beta_2}} G(x_i;\lambda_i) = \partial_{\lambda_i} G(x_i;\lambda_i)\partial_{\boldsymbol{\beta_2}}\lambda_i(x_i;\boldsymbol{\beta_2})$ denotes the derivatives of the parameter λ_i using the chain rule.

The MLE $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ of the regression model is calculated setting the score equations $U_\alpha = U_\theta = U_{\mu_i} = U_{\lambda_i} = 0$ using an iterative method algorithm to find roots or using the *optim* routine.

6.4.2 Simulation study

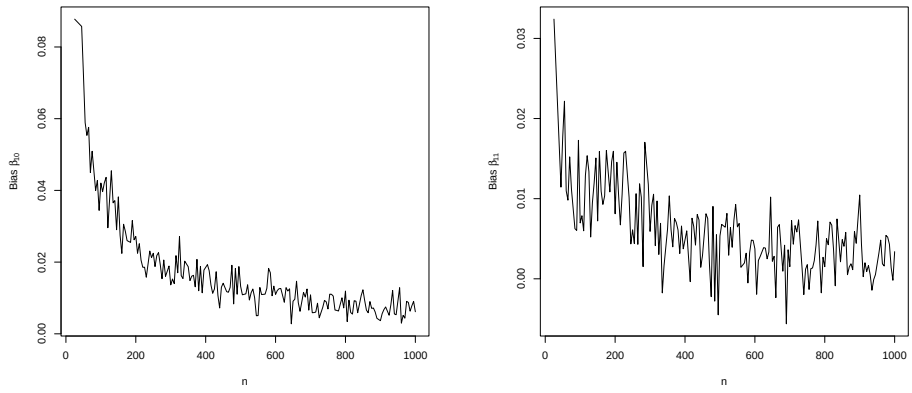
One thousand samples from Equation (6.5) were generated using the parameter vector $(\alpha, \theta, \beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}) = (0.65, 1.25, 0.55, 2.15, 1.44, 0.87)$ for sizes $n = 25, \ldots, 1,000$. For each parameter, the biases, MSEs, ALs and CPs are determined.

Figures 50-59 present these measurements in relation to sample size. The first indicates that, as n increases, the measures converge toward zero, holding the MLEs' consistency criteria. Finally, as the sample size increases, the CPs approach 0.95. The biases of the parameters α and β_{20} are underestimated, nevertheless, the convergence to zero is assured. At last, the parameter β_{21} presents an oscillatory behavior, however, towards zero, as expected.



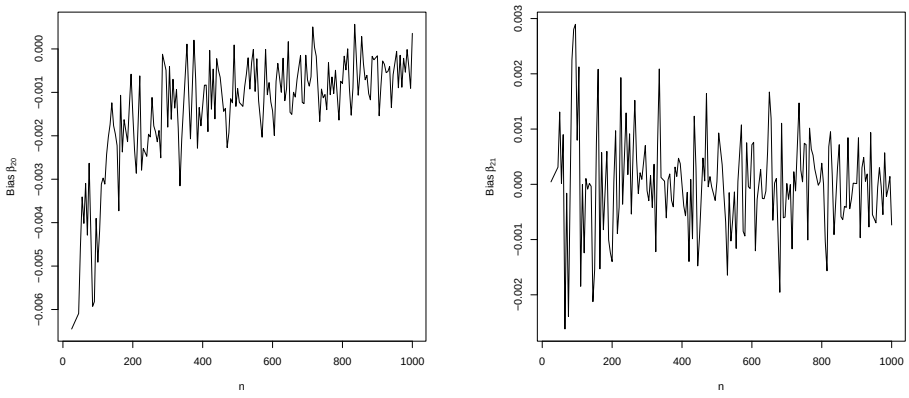
(a) (b)

Figure 50 – Biases versus sample size from GOLLW regression model.



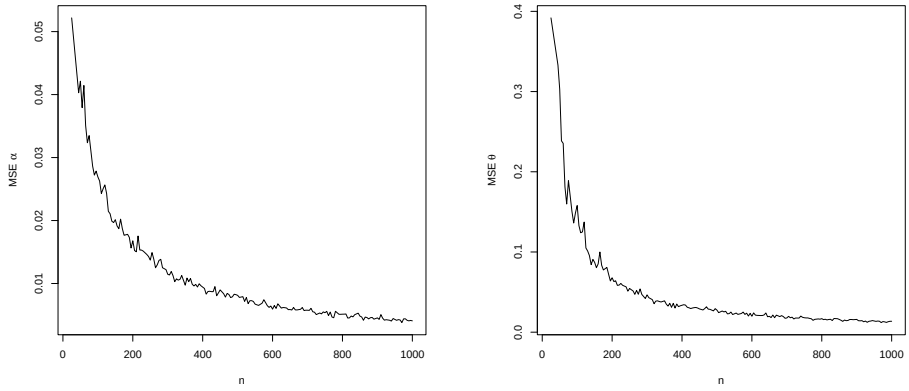
(a) (b)

Figure 51 – Biases versus sample size from GOLLW regression model.



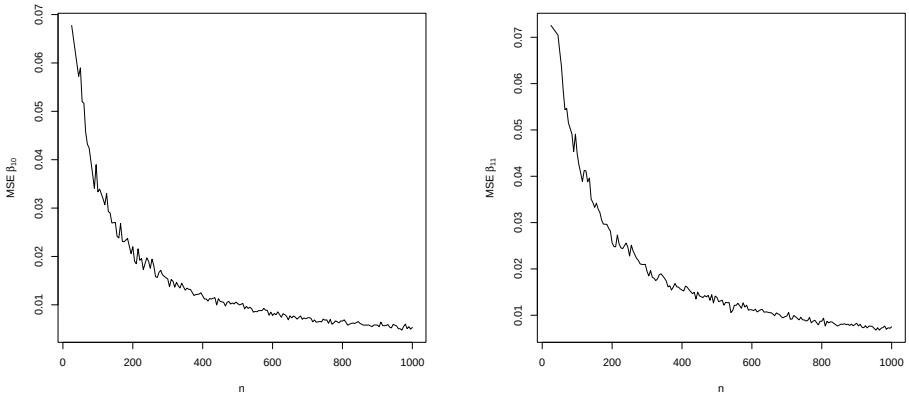
(a) (b)

Figure 52 – Biases versus sample size from GOLLW regression model.



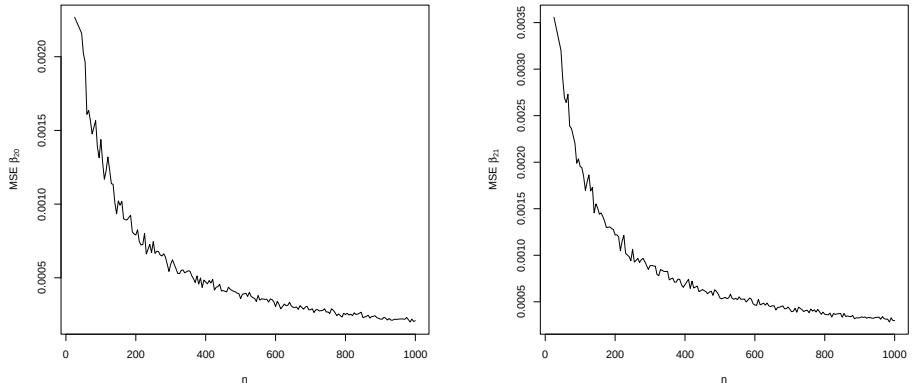
(a) (b)

Figure 53 – Biases versus sample size from GOLLW regression model.



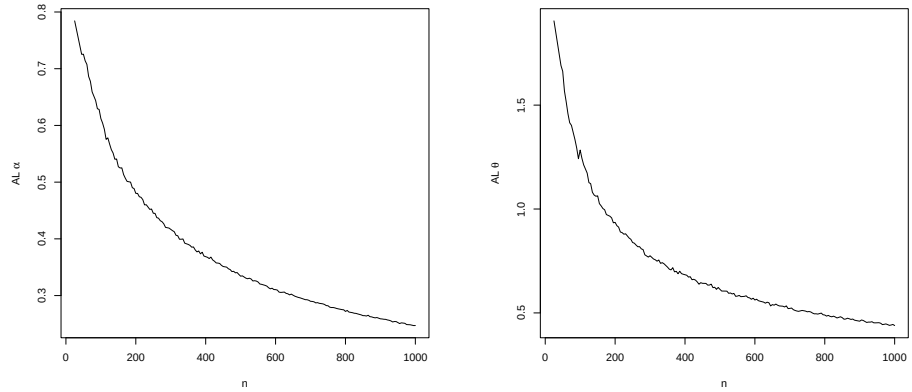
(a) (b)

Figure 54 – Biases versus sample size from GOLLW regression model.



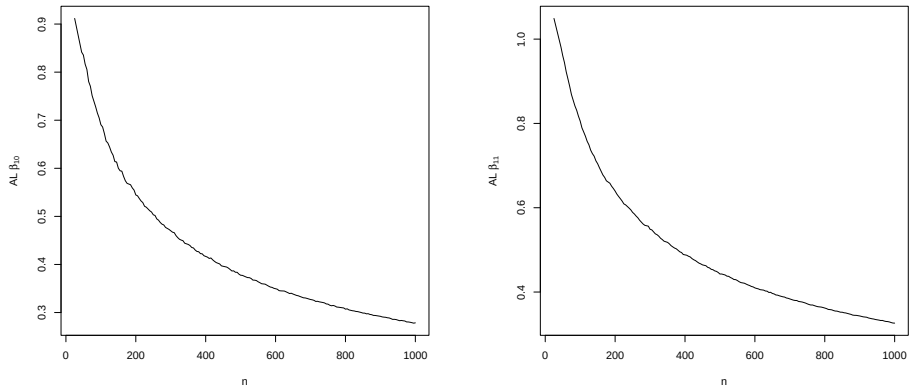
(a) (b)

Figure 55 – Biases versus sample size from GOLLW regression model.



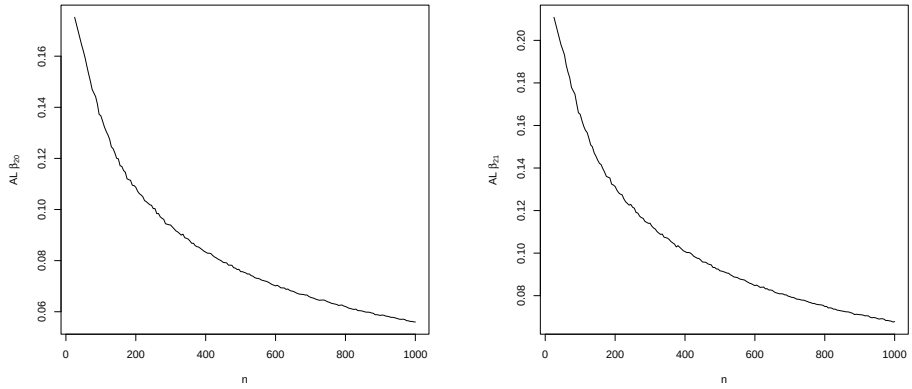
(a) (b)

Figure 56 – Biases versus sample size from GOLLW regression model.



(a) (b)

Figure 57 – Biases versus sample size from GOLLW regression model.



(a) (b)

Figure 58 – Biases versus sample size from GOLLW regression model.

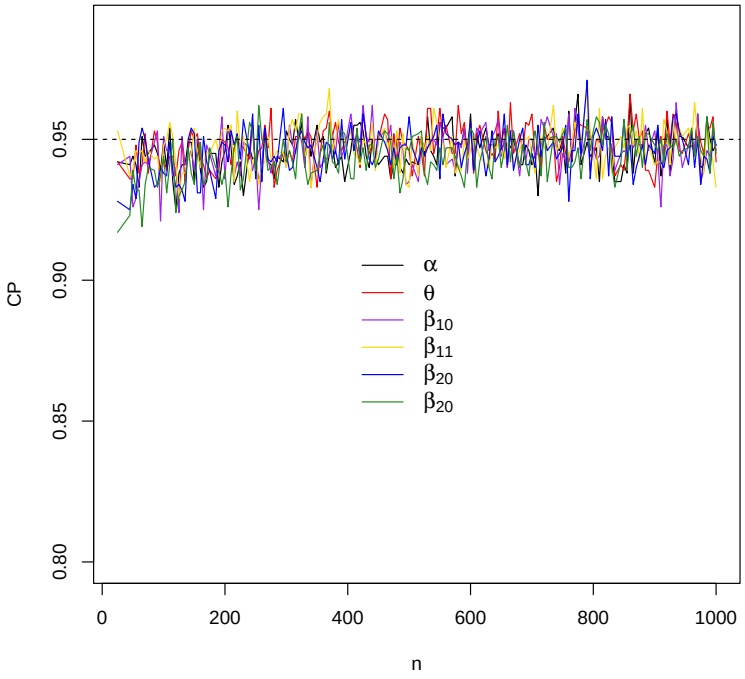


Figure 59 – CPs versus sample size from GOLLW regression model.

6.5 APPLICATION: BRAZIL DAILY WIND ENERGY GENERATION

This Section includes an example application to demonstrate the effectiveness of the GOLLW regression model. In addition to nested models, other non-nested models are fitted to compare the new proposed distribution shown in Table 25.

The distributions are displayed (for $x > 0$), respectively:

Table 25 – Competitive distributions compared to the GOLLW distribution.

Distribution	Reference
Beta-Weibull (BW)	(LEE; FAMOYE; OLUMOLADE, 2007)
Kumaraswamy-Weibull (KwW)	(CORDEIRO; ORTEGA; NADARAJAH, 2010)
Gamma-Weibull (GW)	(PROVOST; SABOOR; AHMAD, 2011)
Kumaraswamy-Fréchet (KwFr)	(MEAD, 2014)

$$F_{\mathsf{BW}}(x) = I_{G(x)}(\beta,\delta) = \frac{1}{B(\beta,\delta)}\int_0^{G(x)} w^{\beta-1}(1-w)^{\delta-1}dw,$$

$$F_{\mathsf{KwW}}(x) = 1-\left\{1-G(x)^{\beta}\right\}^{\delta},$$

$$F_{\mathsf{GW}}(x) = \frac{\gamma\{\beta,-\log[1-G(x)]/\delta\}}{\Gamma(\delta)}$$

and

$$F_{\mathsf{KwFr}}(x) = 1-\left\{1-F_{\mathsf{Fr}}(x)^{\beta}\right\}^{\delta},$$

where all the parameters are positive and $G(x)$ and $F_{\mathsf{Fr}}(x)$ is defined in Equation (6.1) and Fréchet distribution, respectively. Using the BFGS procedure, the *goodness.fit* function determines the MLEs (SEs in parentheses) for all fitted models.

6.5.1 A data set definition

The data set corresponds to $n = 365$ observations of total daily wind energy generation in Brazil from January 1st, 2022 to December 31st, 2022 retrieved from ONS[16]. The major goal is to explore the relationship between the total and dispersion of daily wind generation and the exploratory variable month during this time scenario. The variables listed below are the focus of the investigation:

- y_i: total daily wind generation in Watts (W) (response variable);
- m_{ij}: month (levels: 0 - January to 11 - December). Thus, for $i = 1,\ldots,365$, and $j = 0,\ldots,11$, dummy variables.

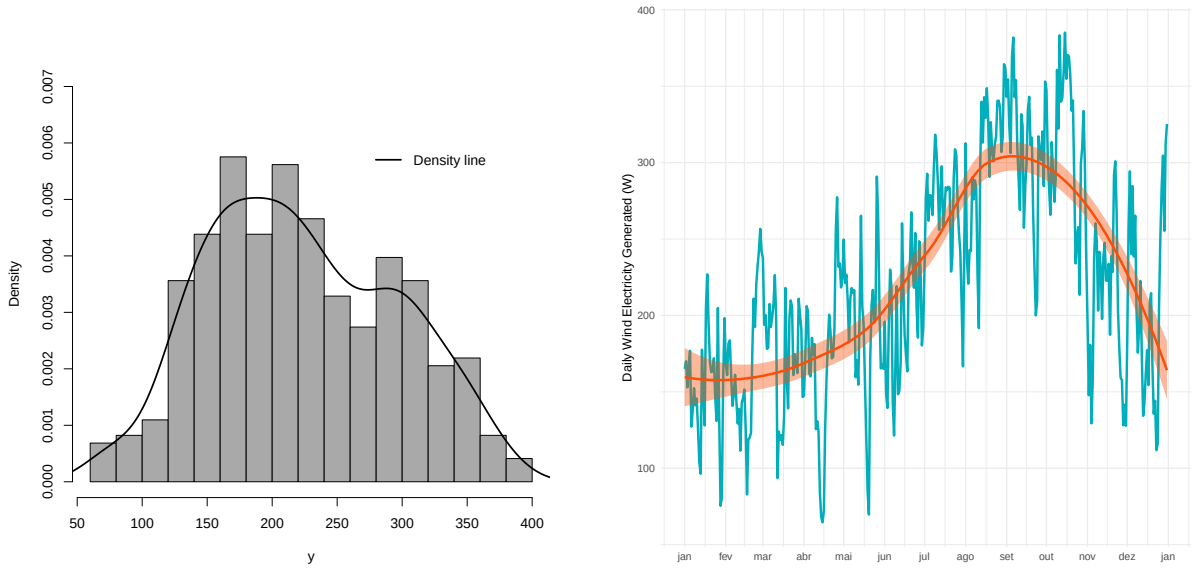
The descriptive statistics of these data, which feature positive asymmetry and kurtosis, are summarized in Table 26. Additionally, the bimodality shape of the data and the aspect of the response variable's throughout the period are displayed in Figure 60(a). In contrast

[16] <https://dados.ons.org.br/dataset/>

to the typical pattern in the remaining months, Figure 60(b) shows the existence of high generation with comparable behavior during the months of July, August, September, and October. Although the wind patterns in different parts of Brazil vary with the seasons, the Northeast is where most energy is generated[17]. Between August and December, there are consistently strong winds in this area, which is indicative of a greater generation trend in August.

Table 26 – Descriptive statistics of Brazil wind energy generation data.

Min.	Max.	Mean	Median	SD	Skewness	Kurtosis
16.34	51.91	33.06	31.04	6.92	0.54	2.48



(a) (b)

Figure 60 – Brazil wind energy generation data. (a) Histogram and empirical density. (b) Variation across months with trend smoothed line.

6.5.2 Results

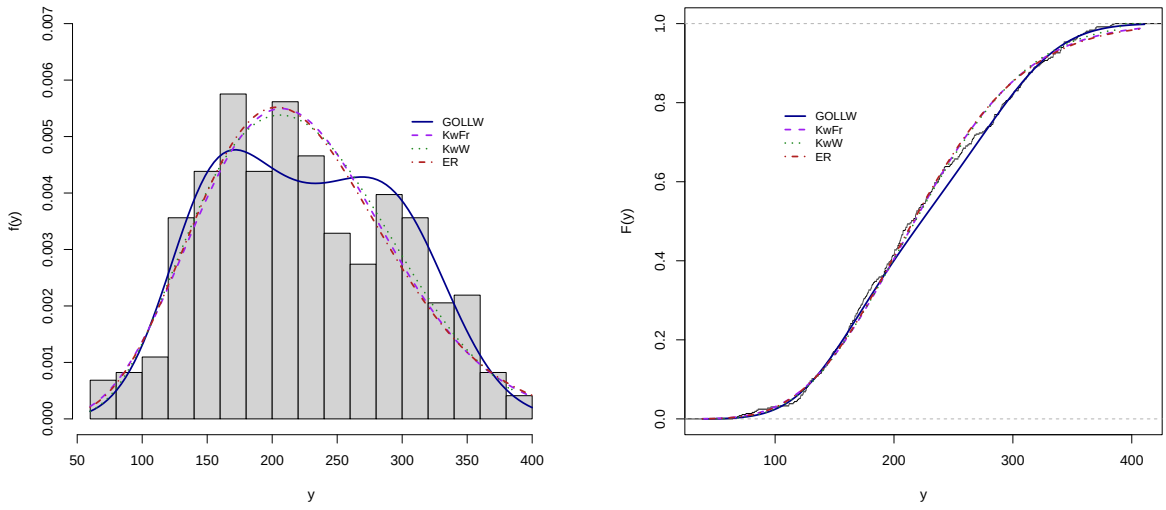
Tables 27 shows the MLEs and GoF measures for nested and non-nested models, respectively and show that the GOLLW distribution offers the best fit to the current data, despite some values of KS measure. In fact, the histogram and plots of the estimated density functions, as well as the empirical and estimated cdf, are shown in Figure 61 to support this result.

As shown in Table 28, the suggested model is highly competitive for both nested and non-nested models. All null hypotheses are rejected by the LR tests in favor of the new distribution.

[17] <https://shorturl.at/fntS8>

Table 27 – Findings from the fitted models of Brazil wind energy data.

Model	α	θ	μ	λ	δ	W*	A*	KS
GOLLW	**0.332**	**4.458**	**3.879**	**188.124**		**0.080**	**0.520**	**0.056**
	(0.082)	**(1.200)**	**(0.496)**	**(10.361)**				**(0.20)**
OLLW	0.903	1	3.711	247.484		0.334	1.938	0.061
	(0.144)	(-)	(0.510)	(4.032)				(0.13)
EW	1	1.468	2.755	222.015		0.224	1.413	0.054
	(-)	(0.495)	(0.505)	(23.905)				(0.24)
W	1	1	3.408	248.052		0.306	1.814	0.056
	(-)	(-)	(0.139)	(4.022)				(0.20)
GOLLE	1.870	3.403	1	126.755		0.243	1.828	0.059
	(0.600)	(2.376)	(-)	(47.937)				(0.16)
OLLE	3.583	1	1	310.807		0.274	1.978	0.061
	(0.155)	(-)	(-)	(5.727)				(0.14)
OLLR	1.763	1	2	261.054		0.251	1.707	0.061
	(0.078)	(-)	(-)	(4.743)				(0.13)
EE	1	15.021	1	67.712		0.251	2.074	0.053
	(-)	(1.777)	(-)	(2.933)				(0.25)
ER	1	2.713	2	177.148		0.161	1.186	0.058
	(-)	(0.225)	(-)	(4.846)				(0.17)
E	1	1	1	222.469		0.150	1.234	0.369
	(-)	(-)	(-)	(11.643)				(<0.01)
R	1	1	2	233.896		0.172	1.183	0.180
	(-)	(-)	(-)	(6.121)				(<0.01)
GW	1.006	-	3.420	248.061		0.308	1.819	0.060
	(0.382)	(-)	(0.808)	(35.231)				(0.15)
BW	1.673	3.853	2.517	343.73		0.213	1.357	0.055
	(0.754)	(18.048)	(0.677)	(587.903)				(0.23)
KwW	2.328	0.116	2.936	113.216		0.117	0.875	0.055
	(0.084)	(0.006)	(0.002)	(0.039)				(0.23)
KwFr	212.135	266.535	-28.952	0.649		0.168	1.196	0.057
	(124.705)	(176.472)	(27.418)	(0.108)				(0.18)



(a) (b)

Figure 61 – Fitted models of Brazil wind energy generation data. (a) Histogram and estimated pdfs. (b) Empirical and estimated cdfs.

Table 28 – LR tests of Brazil wind energy data.

Models	**Statistic w**	**p-value**
GOLLW vs W	16.466	2.657e-04
GOLLW vs EW	15.089	1.025e-04
GOLLW vs OLLW	16.082	6.067e-05
GOLLW vs R	148.829	4.713e-32
GOLLW vs E	539.694	1.191e-116
GOLLW vs ER	18.153	1.143e-04
GOLLW vs EE	43.481	3.616e-10
GOLLW vs OLLR	25.829	2.462e-06
GOLLW vs OLLE	36.439	1.223e-08
GOLLW vs GOLLE	34.718	3.810e-09

Here, based on non-linear equations (for $i = 1, \ldots, 365$), the systematic structures are consider as follows:

$$\mu_i = \exp\left(\beta_{10} + \sum_{j=1}^{11} \beta_{1j}\, m_{ij}\right)$$

and

$$\lambda_i = \exp\left(\beta_{20} + \sum_{j=1}^{11} \beta_{2j}\, m_{ij}\right).$$

The GCD and LD measures, presented in Figure 62, found some influential observations. It's worth noting that the 221st observation (daily wind generation of 9th August 2022 - 283.9285 W) is a potentially influential observation.

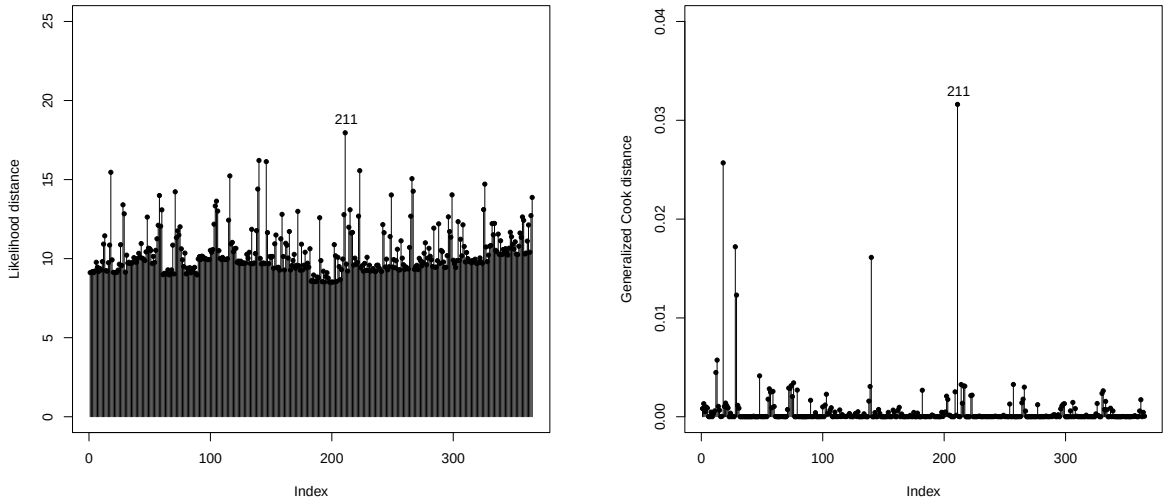
The ONS[18] registry of records in daily wind generation for the entire month of August explains how the day of August 9, 2022, can affect the model, as well as by the extratropical cyclone phenomenon[19], which resulted in a strong wind season in Ceará, the Brazilian state with the highest wind energy production.

Regardless, Figure 63 shows that index deviation residuals have a random behavior in the range and the residuals lie within the simulated envelope, indicating that the observation has no substantial impact on the regression model.

Table 29 presents the findings (MLEs, SEs, and p-values) of the fitted GOLLW heteroscedastic regression model to the current data.

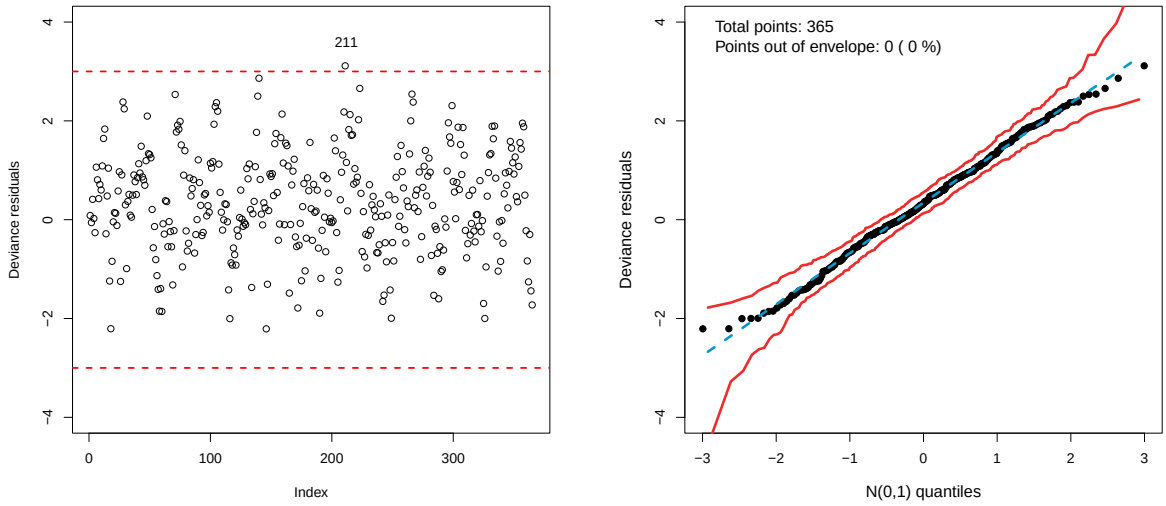
[18] <https://www.ons.org.br/Paginas/Noticias/20220906-Em-agosto,-ONS-registra-31-recordes-na-produ%C3%A7%C3%A3o-de-energia-por-gera%C3%A7%C3%A3o-solar-e-e%C3%B3lica-.aspx>

[19] <https://shorturl.at/enoyW>



(a) (b)

Figure 62 – The GOLLW regression model. (a) LD. (b) GCD.



(a) (b)

Figure 63 – The GOLLW regression model. (a) Deviance residual index. (b) Simulated envelope.

6.5.3 Discussion

The GOLLW regression model is able to describe the wind generation data from Brazil, as demonstrated by the model checks. The GOLLW regression model is shown by using the parameter estimates shown in Table 29 as follows:

Table 29 – Fitted GOLLW regression of Brazil wind energy data.

Parameter	Estimate	SE	p-value	Parameter	Estimate	SE	p-value
$\boldsymbol{\beta_{10}}$	**1.7349**	**0.1742**	$<$ **0.0001**	$\boldsymbol{\beta_{20}}$	**5.0774**	**0.0901**	$<$ **0.0001**
β_{11}	- 0.1924	0.2011	0.3394	β_{21}	0.0921	0.0676	0.1741
β_{12}	0.2044	0.1987	0.3042	$\boldsymbol{\beta_{22}}$	**0.1380**	**0.0533**	**0.0100**
β_{13}	- 0.3048	0.1989	0.1265	β_{23}	0.1038	0.0794	0.1920
β_{14}	- 0.0524	0.1950	0.7883	$\boldsymbol{\beta_{24}}$	**0.2296**	**0.0610**	**0.0002**
β_{15}	0.1789	0.1968	0.3640	$\boldsymbol{\beta_{25}}$	**0.2561**	**0.0555**	$<$ **0.0001**
$\boldsymbol{\beta_{16}}$	**0.8245**	**0.1976**	$<$ **0.0001**	$\boldsymbol{\beta_{26}}$	**0.5468**	**0.0624**	$<$ **0.0001**
$\boldsymbol{\beta_{17}}$	**0.6132**	**0.2001**	**0.0024**	$\boldsymbol{\beta_{27}}$	**0.6631**	**0.0570**	$<$ **0.0001**
$\boldsymbol{\beta_{18}}$	**0.5849**	**0.1984**	**0.0034**	$\boldsymbol{\beta_{28}}$	**0.6857**	**0.0580**	$<$ **0.0001**
$\boldsymbol{\beta_{19}}$	**0.5319**	**0.1972**	**0.0074**	$\boldsymbol{\beta_{29}}$	**0.7082**	**0.0575**	$<$ **0.0001**
β_{110}	0.0062	0.1948	0.9748	$\boldsymbol{\beta_{210}}$	**0.3042**	**0.0584**	$<$ **0.0001**
β_{111}	-0.2192	0.1960	0.2642	$\boldsymbol{\beta_{211}}$	**0.3350**	**0.0670**	$<$ **0.0001**

$$\begin{aligned}\hat{\mu}_i &= \exp(1.7349 + 0.8245\, m_{i6} + 0.6132\, m_{i7} + 0.5849\, m_{i8} + 0.5319\, m_{i9}),\\ \hat{\lambda}_i &= \exp(5.0774 + 0.1380\, m_{i2} + 0.2296\, m_{i4} + 0.2561\, m_{i5} + 0.5468\, m_{i6}\\ &\quad + 0.6631\, m_{i7} + 0.6857\, m_{i8} + 0.7082\, m_{i9} + 0.3042\, m_{i10} + 0.3350\, m_{i11}).\end{aligned} \tag{6.10}$$

The discussion below presents an analysis of the systematic structures, utilizing January as the month reference. In addition, Figure 64 illustrates the estimated cdfs of the fitted GOLLW regression model and the empirical ones in specified months.

- **Interpretations of the systematic component μ**
  - July, August, September, and October (β_{16} to β_{19}) are statistically significant at a 5% confidence level. The positive estimates imply a positive impact on the baseline distribution's shape parameter, which may displace the mean of daily wind generation. The non-significant months are due the similar behavior to the reference month (January);
  - Figure 64 shows the absence of a significant difference among the months of February, March, April, May, June, November and December. The same is visualized in July, and August, September, and October. The new regression detects similar characteristics in the shape and scale parameters, namely clusters, and shows an increase in energy generation in the months of July, August, September and October. Therefore, there are three clusters, namely "July", "August to "October" and the rest of the months;

– The highest wind generation came from the last cluster. In turn, the lowest generation is in the largest cluster.

- **Interpretations of the systematic component λ**

  – With the exception of February (β_{21}) and April (β_{23}), all other covariables exhibit statistical significance at the 5% level. The estimates are positive and evidencing an effect on increasing the scale parameter and, as a result, an increase in kurtosis with a positive impact on the spread of wind energy generation between the other months in relation to January, as noted in Figure 60(b).

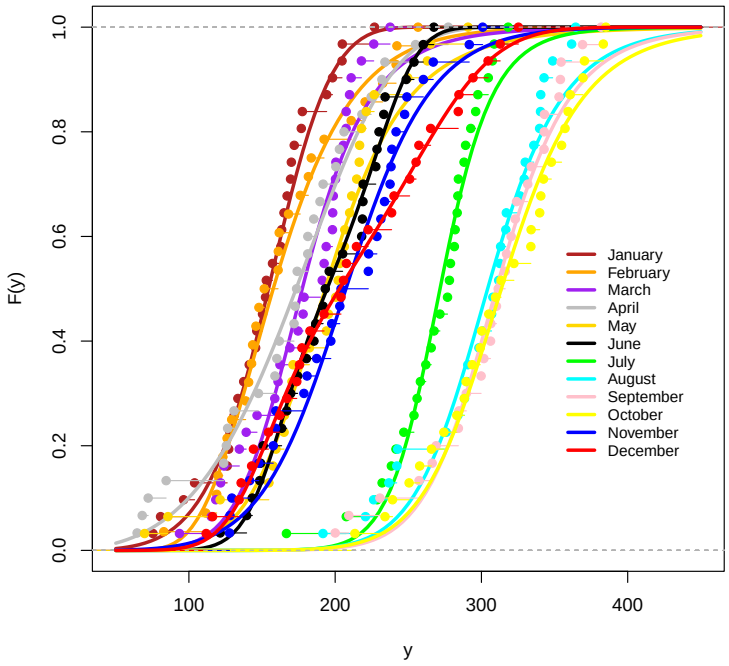


Figure 64 – Estimated and empirical cdfs.

6.6 CONCLUDING REMARKS

The Chapter presented the *generalized odd log-logistic Weibull* distribution (CORDEIRO et al., 2017) and developed a novel Weibull bimodal regression model with two systematic structure to analyze the total daily wind energy generation in Brazil from January 1st, 2022 to December 31st, 2022. Some valuable characteristics are addressed, the parameters are estimated via maximum likelihood and by means of Monte Carlo simulations, some measures

are used for the MLEs. Simulations are used to evaluate the consistency of the regression model's MLEs. Based on influence diagnostics and residual analysis, the performance of the model is investigated.

Some remarkable findings are presented. For the shape systematic component, the months of July to October are significant with positive estimates. This indicates a positive impact on the shape parameter of the baseline, which may shift the mean of daily wind generation. For the scale systematic component, except for the months of February and April, all other months are significant with a positive estimate. This provides an increase in the scale parameter and, as a consequence, in the kurtosis, implying a positive impact on the spread of wind energy generation data.

The applicability of the energy generation data in the proposed regression model showed more usefulness and with a better fit than nested and competitive models. In consequence, the model improves the comprehension of Brazil daily wind energy generation.

7 CONCLUSION

In this work, based on the *generalized odd log-logistic-G* family of distributions proposed by Cordeiro et al. (2017), a few parametric regression models are introduced. Using one- and two-parameter distributions with different regression structures, some densities are presented to investigate the relationship between explanatory variables and the baseline vector parameter, enabling further inspection and interpretation of various phenomena. Linear representations and various mathematical properties are addressed for all the introduced models. The adaptability makes the model a powerful tool for accommodating diverse types of data shapes, which include asymetric, bimodal and fat tails.

Considering one-parameter distributions, the GOLLL and the GOLLE regression models were defined. The first mentioned analyzed the completed primary COVID-19 vaccination rate of counties in the state of Texas, US and effectively identified the factors that influence COVID-19 vaccination rate. The secondary study examines extreme events that impact the response variable. In the Chapter, cases from an epidemiological weekly of dengue fever are used in the Federal District. The results support prior research and indicate that the new models are a suitable alternative to other competitive distributions.

At last, utilizing the well-known distributions gamma and Weibull, the GOLLΓ and GOLLW are defined. Two bimodal regression model structures are presented. The first mentioned illustrates an application to agricultural data of a native plant of the Peruvian Andes, called the yacon potato. The following one was studied by means of a regression model of Brazilian wind energy generation. The GOLLΓ regression model proved to be more accurate than prior studies, demonstrating the versatility of the regression. Moreover, the GOLLW regression provided strong support to analyze and identify elements that impact wind generation.

As a future perspective for this work, the following alternatives can be highlighted: based on all the GOLL-G models proposed, consider a semiparametric regression model, usage of cure rate, additive partial linear regression and inflated zero models can be utilized to widen the model's application, as well as more analysis of the systematic structure and its interpretability to be more accurate.

REFERENCES

ADEPOJU, K.; CHUKWU, O. Maximum likelihood estimation of the Kumaraswamy exponential distribution with applications. *Journal of Modern Applied Statistical Methods*, v. 14, n. 1, p. 18, 2015.

AFIFY, A. Z.; ALIZADEH, M.; ZAYED, M.; RAMIRES, T. G.; LOUZADA, F. The odd log-logistic exponentiated Weibull distribution: Regression modeling, properties, and applications. *Iranian Journal of Science and Technology, Transactions A: Science*, Springer, v. 42, p. 2273–2288, 2018.

AFIFY, A. Z.; SUZUKI, A. K.; ZHANG, C.; NASSAR, M. On three-parameter exponential distribution: Properties, Bayesian and non-Bayesian estimation based on complete and censored samples. *Communications in Statistics - Simulation and Computation*, Taylor Francis, v. 50, p. 3799–3819, 2021.

AGARWAL, R.; DUGAS, M.; RAMAPRASAD, J.; LUO, J.; LI, G.; GAO, G. Socioeconomic privilege and political ideology are associated with racial disparity in COVID-19 vaccination. *Proceedings of the National Academy of Sciences*, v. 118, n. 33, p. e2107873118, 2021.

ALBRECHT, D. Vaccination, politics and COVID-19 impacts. *BMC Public Health*, BioMed Central, v. 22, n. 1, p. 1–12, 2022.

ALIZADEH, M.; LAK, F.; RASEKHI, M.; RAMIRES, T. G.; YOUSOF, H. M.; ALTUN, E. The odd log-logistic Topp-Leone G family of distributions: Heteroscedastic regression models and applications. *Computational Statistics*, Springer, v. 33, p. 1217–1244, 2018.

ALTUN, E.; ALIZADEH, M.; OZEL, G.; YOUSOF, H. M. New odd log-logistic family of distributions: Properties, regression models and applications. In: *G Families of Probability Distributions*. [S.l.]: CRC Press, 2023. p. 80–93.

ALTUN, E.; ALIZADEH, M.; YOUSOF, H. M.; RASEKHI, M.; HAMEDANI, G. A new type II half logistic-G family of distributions with properties, regression models, system reliability and applications. *Applications and Applied Mathematics: An International Journal (AAM)*, v. 16, n. 2, p. 3, 2021.

ALTUN, E.; ALIZADEH, M.; YOUSOF, H. M.; HAMEDANI, G. The Gudermannian generated family of distributions with characterizations, regression models and applications. *Studia Scientiarum Mathematicarum Hungarica*, Akadémiai Kiadó Budapest, v. 59, n. 2, p. 93–115, 2022.

ALZAATREH, A.; LEE, C.; FAMOYE, F. A new method for generating families of continuous distributions. *METRON*, v. 71, 2013.

ASGHARZADEH, A.; NADARAJAH, S.; SHARAFI, F. Weibull Lindley distribution. *Revstat - Statistical Journal*, v. 16, 01 2016.

ATKINSON, A. *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. [S.l.]: Clarendon Press, 1987. (Oxford science publications).

BARRY, V.; DASGUPTA, S.; WELLER, D. L.; KRISS, J. L.; CADWELL, B. L.; ROSE, C.; PINGALI, C.; MUSIAL, T.; SHARPE, J. D.; FLORES, S. A. et al. Patterns in COVID-19 vaccination coverage, by social vulnerability and urbanicity — United States, December 14, 2020 - May 1, 2021. *Morbidity and Mortality Weekly Report*, Centers for Disease Control and Prevention, v. 70, n. 22, p. 818, 2021.

BERRY, D. M.; ADAMS, L. M.; VYTLA, S. P. Vaccine hesitancy and hesitant adoption among nursing students in Texas. *Preventive Medicine Reports*, Elsevier, v. 38, p. 102612, 2024.

BUHEJI, M.; CUNHA, K. da C.; BEKA, G.; MAVRIC, B.; SOUZA, Y. D.; SILVA, S. S. da C.; HANAFI, M.; YEIN, T. C. The extent of COVID-19 pandemic socio-economic impact on global poverty. A global integrative multidisciplinary review. *American Journal of Economics*, Scientific and Academic Publishing, v. 10, n. 4, p. 213–224, 2020.

ÇAKMAKYAPAN, S.; KADILAR, G. Ö. A new customer lifetime duration distribution: The Kumaraswamy Lindley distribution. *International Journal of Trade, Economics and Finance*, IACSIT Press, v. 5, n. 5, p. 441, 2014.

CHOI, W. S.; CHEONG, H. J. COVID-19 vaccination for people with comorbidities. *Infection & Chemotherapy*, Korean Society of Infectious Diseases, v. 53, n. 1, p. 155, 2021.

CHOUDHURY, A. A simple derivation of moments of the exponentiated Weibull distribution. *Metrika*, Springer, v. 62, p. 17–22, 2005.

CLERKIN, K. J.; FRIED, J. A.; RAIKHELKAR, J.; SAYER, G.; GRIFFIN, J. M.; MASOUMI, A.; JAIN, S. S.; BURKHOFF, D.; KUMARAIAH, D.; RABBANI, L. et al. COVID-19 and cardiovascular disease. *Circulation*, Am Heart Assoc, v. 141, n. 20, p. 1648–1655, 2020.

COOK, R.; WEISBERG, S. *Residuals and Influence in Regression*. [S.l.]: Taylor & Francis, 1982. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability).

COORAY, K. Generalization of the Weibull distribution: The odd Weibull family. *Statistical Modelling*, Sage Publications Sage CA: Thousand Oaks, CA, v. 6, n. 3, p. 265–277, 2006.

CORDEIRO, G. M.; ALIZADEH, M.; OZEL, G.; HOSSEINI, B.; ORTEGA, E. M. M.; ALTUN, E. The generalized odd log-logistic family of distributions: Properties, regression models and applications. *Journal of statistical computation and simulation*, Taylor & Francis, v. 87, n. 5, p. 908–932, 2017.

CORDEIRO, G. M.; CASTRO, M. de. A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 81, n. 7, p. 883–898, 2011.

CORDEIRO, G. M.; ORTEGA, E. M.; NADARAJAH, S. The Kumaraswamy Weibull distribution with application to failure data. *Journal of the Franklin Institute*, Elsevier, v. 347, n. 8, p. 1399–1429, 2010.

CORDEIRO, G. M.; RODRIGUES, G. M.; ORTEGA, E. M.; SANTANA, L. H. de; VILA, R. An extended Rayleigh model: Properties, regression and COVID-19 application. *arXiv preprint arXiv:2204.05214*, 2022.

COŞKUN, K.; KORKMAZ, M. Ç.; KINACI, İ.; KARAKAYA, K.; AKDOĞAN, Y. Modified-Lindley distribution and its applications to the real data. *Communications Faculty of Sciences University of Ankara Series A1 Mathematics and Statistics*, Ankara University, v. 71, n. 1, p. 252–272, 2022.

COUGHENOUR, C.; GAKH, M.; SHARMA, M.; LABUS, B.; CHIEN, L.-C. Assessing determinants of COVID-19 vaccine hesitancy in Nevada. *Health security*, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New . . . , v. 19, n. 6, p. 592–604, 2021.

COX, D. R.; SNELL, E. J. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 30, p. 248–275, 1968.

CRUZ, J. N. d.; ORTEGA, E. M.; CORDEIRO, G. M. The log-odd log-logistic Weibull regression model: Modelling, estimation, influence diagnostics and residual analysis. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 86, n. 8, p. 1516–1538, 2016.

DEATON, A. *COVID-19 and Global Income Inequality*. [S.l.], 2021. (Working Paper Series, 28392).

DELINA, L. L. Wind energy in the city: Hong Kong's offshore wind energy generation potential, deployment plans, and ecological pitfalls. *The Electricity Journal*, Elsevier, v. 35, n. 6, p. 107139, 2022.

DIOP, A.; DEME, E. H.; DIOP, A. Zero-inflated generalized extreme value regression model for binary data and application in health study. *arXiv preprint arXiv:2105.00482*, 2021.

DOHERTY, I. A.; PILKINGTON, W.; BROWN, L.; BILLINGS, V.; HOFFLER, U.; PAULIN, L.; KIMBRO, K. S.; BAKER, B.; ZHANG, T.; LOCKLEAR, T. et al. COVID-19 vaccine hesitancy in underserved communities of North Carolina. *PLoS One*, Public Library of Science San Francisco, CA USA, v. 16, n. 11, p. e0248542, 2021.

ELGARHY, M.; MUTAIRI, A. A.; HASSAN, A. S.; CHESNEAU, C.; ABDEL-HAMID, A. H. Bayesian and non-Bayesian estimations of truncated inverse power Lindley distribution under progressively type-II censored data with applications. *AIP Advances*, AIP Publishing, v. 13, n. 9, 2023.

Felipe de Mendiburu; Muhammad Yaseen. *agricolae: Statistical Procedures for Agricultural Research*. [S.l.], 2020. R package version 1.4.0.

FRÉCHET, M. Sur la loi de probabilité de l'écart maximum. *Ann. de la Soc. Polonaise de Math.*, 1927.

GALTON, F. *Inquiries into Human Faculty and its Development*. [S.l.]: Macmillan, 1883.

GLEATON, J.; LYNCH, J. Properties of generalized log-logistic families of lifetime distributions. *Journal of Probability and Statistical Science*, v. 4, p. 51–64, 01 2006.

GOEL, R. K.; NELSON, M. A. COVID-19 internet vaccination information and vaccine administration: Evidence from the United States. *Journal of Economics and Finance*, Springer, v. 45, n. 4, p. 716–734, 2021.

GRADSHTEYN, I.; JEFFREY, A.; RYZHIK, I. *Table of Integrals, Series, and Products*. [S.l.]: Academic Press, 1996.

GUAN, W.-j.; LIANG, W.-h.; ZHAO, Y.; LIANG, H.-r.; CHEN, Z.-s.; LI, Y.-m.; LIU, X.-q.; CHEN, R.-c.; TANG, C.-l.; WANG, T. et al. Comorbidity and its impact on 1,590 patients with COVID-19 in China: A nationwide analysis. *European Respiratory Journal*, Eur Respiratory Soc, v. 55, n. 5, 2020.

GUPTA, R.; GUPTA, R. Proportional reversed hazard rate model and its applications. *Journal of Statistical Planning and Inference*, v. 137, p. 3525–3536, 2007.

GUPTA, R. D.; KUNDU, D. Theory & Methods: Generalized Exponential Distributions. *Australian & New Zealand Journal of Statistics*, Wiley Online Library, v. 41, n. 2, p. 173–188, 1999.

GUPTA, R. D.; KUNDU, D. Exponentiated exponential family: An alternative to gamma and Weibull distributions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, Wiley Online Library, v. 43, n. 1, p. 117–130, 2001.

HAMEED, A. N.; SAIEED, H. A. J. Construction of Marshall-Olkin Lindley distribution with application. In: IEEE. *2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM)*. [S.l.], 2022. p. 437–441.

HOSEK, M. G.; CHIDESTER, A. B.; GELFOND, J.; TAYLOR, B. S. Low prevalence of COVID-19 vaccine hesitancy in students across health science disciplines in Texas. *Vaccine: X*, Elsevier, v. 10, p. 100154, 2022.

HUGHES, M. M.; WANG, A.; GROSSMAN, M. K.; PUN, E.; WHITEMAN, A.; DENG, L.; HALLISEY, E.; SHARPE, J. D.; USSERY, E. N.; STOKLEY, S. et al. County-level COVID-19 vaccination coverage and social vulnerability — United States, December 14, 2020 - March 1, 2021. *Morbidity and Mortality Weekly Report*, Centers for Disease Control and Prevention, v. 70, n. 12, p. 431, 2021.

HYDER, A. A.; HYDER, M. A.; NASIR, K.; NDEBELE, P. Inequitable COVID-19 vaccine distribution and its effects. *Bulletin of the World Health Organization*, World Health Organization, v. 99, n. 6, p. 406, 2021.

IBRAHIM, M. The generalized odd log-logistic Nadarajah Haghighi distribution: Statistical properties and different methods of estimation. v. 15, p. 61–84, 08 2020.

ISHAQ, A. I.; ABIODUN, A. A. The Maxwell-Weibull distribution in modeling lifetime datasets. *Annals of Data Science*, Springer, v. 7, n. 4, p. 639–662, 2020.

KARAKAYA, K.; KORKMAZ, M.; CHESNEAU, C.; HAMEDANI, G. A new alternative unit-Lindley distribution with increasing failure rate. *Scientia Iranica*, Sharif University of Technology, 2022.

KHAIRAT, S.; ZOU, B.; ADLER-MILSTEIN, J. Factors and reasons associated with low COVID-19 vaccine uptake among highly hesitant communities in the US. *American Journal of Infection Control*, Elsevier, v. 50, n. 3, p. 262–267, 2022.

KHALEQ, R. H. A. E. The generalized odd log-logistic Fréchet distribution for modeling extreme values. *Pakistan Journal of Statistics and Operation Research*, College of Statistical and Actuarial Sciences, p. 649–674, 2022.

KHAMEES, A. K.; ABDELAZIZ, A. Y.; ALI, Z. M.; ALHARTHI, M. M.; GHONEIM, S. S.; ESKAROS, M. R.; ATTIA, M. A. Mixture probability distribution functions using novel metaheuristic method in wind speed modeling. *Ain Shams Engineering Journal*, Elsevier, v. 13, n. 3, p. 101613, 2022.

KLAKATTAWI, H. S. The Weibull-gamma distribution: Properties and applications. *Entropy*, MDPI, v. 21, n. 5, p. 438, 2019.

KRISS, J. L.; HUNG, M.-C.; SRIVASTAV, A.; BLACK, C. L.; LINDLEY, M. C.; LEE, J. T.; KOPPAKA, R.; TSAI, Y.; LU, P.-J.; YANKEY, D. et al. COVID-19 vaccination coverage, by race and ethnicity — national immunization survey adult COVID module, United States, December 2020 - November 2021. *Morbidity and Mortality Weekly Report*, Centers for Disease Control and Prevention, v. 71, n. 23, p. 757, 2022.

KUDRIAVTSEV, A. A. On the representation of gamma-exponential and generalized negative binomial distributions. *Informatika i Ee Primeneniya [Informatics and its Applications]*, Russian Academy of Sciences, Branch of Informatics, Computer Equipment and . . . , v. 13, n. 4, p. 76–80, 2019.

LAWLESS, J. F. Inference in the generalized gamma and log gamma distributions. *Technometrics*, Taylor & Francis, v. 22, n. 3, p. 409–419, 1980.

LEE, C.; FAMOYE, F.; OLUMOLADE, O. Beta-Weibull distribution: Some properties and applications to censored data. *Journal of Modern Applied Statistical Methods*, v. 6, n. 1, p. 17, 2007.

LI, Y.; DOU, Q.; LU, Y.; XIANG, H.; YU, X.; LIU, S. Effects of ambient temperature and precipitation on the risk of dengue fever: A systematic review and updated meta-analysis. *Environmental Research*, Elsevier, v. 191, p. 110043, 2020.

LIAO, T. F. Social and economic inequality in coronavirus disease 2019 vaccination coverage across Illinois counties. *Scientific reports*, Nature Publishing Group UK London, v. 11, n. 1, p. 18443, 2021.

LIM, J. T.; DICKENS, B. S. L.; COOK, A. R. Modelling the epidemic extremities of dengue transmissions in Thailand. *Epidemics*, Elsevier, v. 33, p. 100402, 2020.

LIN, H.; ZHANG, Z. Extreme co-movements between infectious disease events and crude oil futures prices: From extreme value analysis perspective. *Energy Economics*, Elsevier, v. 110, p. 106054, 2022.

LUN, X.; WANG, Y.; ZHAO, C.; WU, H.; ZHU, C.; MA, D.; XU, M.; WANG, J.; LIU, Q.; XU, L. et al. Epidemiological characteristics and temporal-spatial analysis of overseas imported dengue fever cases in outbreak provinces of China, 2005-2019. *Infectious Diseases of Poverty*, BioMed Central, v. 11, n. 1, p. 1–17, 2022.

LUNINGHAM, J. M.; AKPAN, I. N.; TASKIN, T.; ALKHATIB, S.; VISHWANATHA, J. K.; THOMPSON, E. L. Demographic and psychosocial correlates of COVID-19 vaccination status among a statewide sample in Texas. *Vaccines*, MDPI, v. 11, n. 4, p. 848, 2023.

MALIK, A. A.; MCFADDEN, S. M.; ELHARAKE, J.; OMER, S. B. Determinants of COVID-19 vaccine acceptance in the US. *EClinicalMedicine*, Elsevier, v. 26, 2020.

MARANI, M.; KATUL, G. G.; PAN, W. K.; PAROLARI, A. J. Intensity and frequency of extreme novel epidemics. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 118, n. 35, p. e2105482118, 2021.

MARCIANO, F. W.; NASCIMENTO, A.; SANTOS-NETO, M.; CORDEIRO, G. The Mc-gamma distribution and its statistical properties: An application to reliability data. *International Journal of Statistics and Probability*, Canadian Center of Science and Education, v. 1, n. 1, p. 53, 2012.

MARINHO, P.; SILVA, R.; BOURGUIGNON, M.; CORDEIRO, G.; NADARAJAH, S. Adequacymodel: An R package for probability distributions and general purpose optimization. *PLOS ONE*, v. 14, p. e0221487, 2019.

MEAD, M. E. A. A note on Kumaraswamy Fréchet distribution. *Australia*, v. 8, p. 294–300, 2014.

MEROVCI, F.; SHARMA, V. K. et al. The beta-Lindley distribution: Properties and applications. *Journal of Applied Mathematics*, Hindawi, v. 2014, 2014.

MICHAELS, I. H.; PIRANI, S. J.; CARRASCAL, A. Peer reviewed: Disparities in internet access and COVID-19 vaccination in New York city. *Preventing Chronic Disease*, Centers for Disease Control and Prevention, v. 18, 2021.

MOFLEH, D.; ALMOHAMAD, M.; OSAGHAE, I.; BEMPAH, S.; ZHANG, Q.; TORTOLERO, G.; EBEIDAT, A.; RAMPHUL, R.; SHARMA, S. V. Spatial patterns of COVID-19 vaccination coverage by social vulnerability index and designated COVID-19 vaccine sites in Texas. *Vaccines*, MDPI, v. 10, n. 4, p. 574, 2022.

MOHAMED, H. S.; ALI, M. M.; YOUSOF, H. M. The Lindley Gompertz model for estimating the survival rates: Properties and applications in insurance. *Annals of Data Science*, Springer, v. 10, n. 5, p. 1199–1216, 2023.

MOORS, J. A quantile alternative for kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, Wiley Online Library, v. 37, n. 1, p. 25–32, 1988.

MSIGWA, G.; IGHALO, J. O.; YAP, P.-S. Considerations on environmental, economic, and energy impacts of wind energy generation: Projections towards sustainability initiatives. *Science of The Total Environment*, Elsevier, v. 849, p. 157755, 2022.

MUDHOLKAR, G. S.; SRIVASTAVA, D. K. Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, IEEE, v. 42, n. 2, p. 299–302, 1993.

MUDHOLKAR, G. S.; SRIVASTAVA, D. K. Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE transactions on reliability*, IEEE, v. 42, n. 2, p. 299–302, 1993.

MURTHY, B. P.; STERRETT, N.; WELLER, D.; ZELL, E.; REYNOLDS, L.; TOBLIN, R. L.; MURTHY, N.; KRISS, J.; ROSE, C.; CADWELL, B. et al. Disparities in COVID-19 vaccination coverage between urban and rural counties— United States, December 14, 2020 - April 10, 2021. *Morbidity and Mortality Weekly Report*, Centers for Disease Control and Prevention, v. 70, n. 20, p. 759, 2021.

NADARAJAH, S. The exponentiated Gumbel distribution with climate application. *Environmetrics*, v. 17, 2006.

NADARAJAH, S.; BAKOUCH, H. S.; TAHMASBI, R. A generalized Lindley distribution. *Sankhya B*, Springer, v. 73, p. 331–359, 2011.

NADARAJAH, S.; GUPTA, A. K. The exponentiated gamma distribution with application to drought data. *Calcutta Statistical Association Bulletin*, SAGE Publications Sage India: New Delhi, India, v. 59, n. 1-2, p. 29–54, 2007.

NADARAJAH, S.; KOTZ, S. The beta exponential distribution. *Reliability Engineering System Safety*, v. 91, n. 6, p. 689–697, 2006. ISSN 0951-8320.

NADARAJAH, S.; KOTZ, S. The exponentiated type distributions. *Acta Applicandae Mathematica*, Springer, v. 92, p. 97–111, 2006.

OLIVEIRA, B. R. B. de; SOBRAL, A. I. G. da P.; MARINHO, M. L. M.; SOBRAL, M. F. F.; MELO, A. de S.; DUARTE, G. B. Determinants of access to the SARS-CoV-2 vaccine: A preliminary approach. *International Journal for Equity in Health*, BioMed Central, v. 20, n. 1, p. 1–11, 2021.

OLIVEIRA-JÚNIOR, J. F. de; SOUZA, A. d.; ABREU, M. C.; NUNES, R. S. C.; NASCIMENTO, L. d. S.; SILVA, S. D. d.; FILHO, W. L. F. C.; SILVA, E. B. d. Modeling of dengue by cluster analysis and probability distribution functions in the State of Alagoas in Brazilian. *Brazilian Archives of Biology and Technology*, SciELO Brasil, v. 66, p. e23220086, 2023.

ORTEGA, E. M.; CRUZ, J. N. da; CORDEIRO, G. M. The log-odd logistic-Weibull regression model under informative censoring. *Model Assisted Statistics and Applications*, IOS Press, v. 14, n. 3, p. 239–254, 2019.

ORTEGA, E. M. M.; PAULA, G. A.; BOLFARINE, H. Deviance residuals in generalised log-gamma regression models with censored observations. *Journal of Statistical Computation and Simulation*, v. 78, p. 747–764, 2008.

OSUAGWU, U. L.; LANGSI, R.; OVENSERI-OGBOMO, G.; MASHIGE, K. P.; ABU, E. K.; ENVULADU, E. A.; GOSON, P. C.; EKPENYONG, B. N.; OLORUNTOBA, R.; MINER, C. A. et al. Analysis of perception, reasons, and motivations for COVID-19 vaccination in people with diabetes across sub-saharan Africa: A mixed-method approach. *International Journal of Environmental Research and Public Health*, MDPI, v. 19, n. 13, p. 7875, 2022.

OZEL MORAD ALIZADEH, S. C. G. G. H. E. M. M. O. G.; CANCHO, V. G. The odd log-logistic Lindley Poisson model for lifetime data. *Communications in Statistics - Simulation and Computation*, Taylor Francis, v. 46, n. 8, p. 6513–6537, 2017.

PAROLIN, Z.; LEE, E. K. The role of poverty and racial discrimination in exacerbating the health consequences of COVID-19. *The Lancet Regional Health - Americas*, v. 7, p. 100178, 2022. ISSN 2667-193X.

PRATAVIERA, F.; CORDEIRO, G. M.; ORTEGA, E. M.; SUZUKI, A. K. The odd log-logistic geometric normal regression model with applications. *Advances in Data Science and Adaptive Analysis*, World Scientific, v. 11, n. 01n02, p. 1950003, 2019.

PRATAVIERA, F.; CORDEIRO, G. M.; ORTEGA, E. M. M.; HASHIMOTO, E. M.; CANCHO, V. G. A new regression model for rates and proportions data with applications. *Journal of Applied Statistics*, Taylor & Francis, v. 49, n. 16, p. 4137–4161, 2022.

PRATAVIERA, F.; HASHIMOTO, E. M.; ORTEGA, E. M. M.; CORDEIRO, G. M.; CANCHO, V. G.; VILA, R. A new flexible regression model with application to recovery probability COVID-19 patients. *Journal of Applied Statistics*, Taylor & Francis, p. 1–19, 2023.

PRATAVIERA, F.; ORTEGA, E.; CORDEIRO, G. A new bimodal Maxwell regression model with engineering applications. *Applied Mathematics & Information Sciences*, v. 14, p. 817–31, 2020.

PRATAVIERA, F.; ORTEGA, E. M.; CORDEIRO, G. M.; PESCIM, R. R.; VERSSANI, B. A. A new generalized odd log-logistic flexible Weibull regression model with applications in repairable systems. *Reliability Engineering System Safety*, v. 176, p. 13–26, 2018.

PRATAVIERA, F.; ORTEGA, E. M.; CORDEIRO, G. M.; BRAGA, A. d. S. The heteroscedastic odd log-logistic generalized gamma regression model for censored data. *Communications in Statistics-Simulation and Computation*, Taylor & Francis, v. 48, n. 6, p. 1815–1839, 2019.

PRATAVIERA, F.; SILVA, A.; CARDOSO, E.; CORDEIRO, G.; ORTEGA, E. A novel generalized odd log-logistic Maxwell-based regression with application to microbiology. *Applied Mathematical Modelling*, v. 93, 2020.

PROVOST, S.; SABOOR, A.; AHMAD, M. The gamma-Weibull distribution. *Pakistan Journal of Statistics*, v. 27, p. 111–131, 04 2011.

QOSHJA, A.; MUÇA, M. A new modified generalized odd log-logistic distribution with three parameters. *Mathematical Theory and Modeling*, v. 8, n. 1, p. 2224–5804, 2018.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>.

RANJBAR, V.; ALIZADEH, M.; ALTUN, E. Extended generalized Lindley distribution: Properties and applications. *Journal of Mathematical Extension*, v. 13, p. 117–142, 2019.

RAZZAGHI, H.; MEGHANI, M.; PINGALI, C.; CRANE, B.; NALEWAY, A.; WEINTRAUB, E.; TAT'YANA, A. K.; LAMIAS, M. J.; IRVING, S. A.; KAUFFMAN, T. L. et al. COVID-19 vaccination coverage among pregnant women during pregnancy — eight integrated health care organizations, United States, December 14, 2020 - May 8, 2021. *Morbidity and Mortality Weekly Report*, Centers for Disease Control and Prevention, v. 70, n. 24, p. 895, 2021.

REIMER, N. K.; ATARI, M.; KARIMI-MALEKABADI, F.; TRAGER, J.; KENNEDY, B.; GRAHAM, J.; DEHGHANI, M. Moral values predict county-level COVID-19 vaccination rates in the United States. *American Psychologist*, American Psychological Association, v. 77, n. 6, p. 743, 2022.

ROGA, S.; BARDHAN, S.; KUMAR, Y.; DUBEY, S. K. Recent technology and challenges of wind energy generation: A review. *Sustainable Energy Technologies and Assessments*, Elsevier, v. 52, p. 102239, 2022.

SANDEEP, M.; PADHI, B. K.; YELLA, S. S. T.; SRUTHI, K.; VENKATESAN, R. G.; SASANKA, K. B. K.; SATAPATHY, P.; MOHANTY, A.; AL-TAWFIQ, J. A.; IQHRAMMULLAH, M. et al. Myocarditis manifestations in dengue cases: A systematic review and meta-analysis. *Journal of Infection and Public Health*, Elsevier, v. 16, n. 11, p. 1761–1768, 2023.

SAYIBU, S. B.; LUGUTERAH, A.; NASIRU, S. Mcdonald generalized power Weibull distribution: Properties, and applications. *J. Stat. Appl. Probab*, v. 13, p. 297–322, 2024.

SHIRKE, D. T.; S., K. C. Ton exponentiated log-normal distribution. *International Journal of Agricultural and Statistical Science*, v. 2, p. 319–326, 2006.

SILVA, G. O.; ORTEGA, E. M. M.; PAULA, G. A. Residuals for log-Burr XII regression models in survival analysis. *Journal of applied statistics*, Taylor & Francis, v. 38, n. 7, p. 1435–1445, 2011.

STACY, E. W. A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, JSTOR, p. 1187–1192, 1962.

STACY, E. W.; MIHRAM, G. A. Parameter estimation for a generalized gamma distribution. *Technometrics*, Taylor & Francis, v. 7, n. 3, p. 349–358, 1965.

SUN, Y.; MONNAT, S. M. Rural-urban and within-rural differences in COVID-19 vaccination rates. *The Journal of Rural Health*, Wiley Online Library, v. 38, n. 4, p. 916–922, 2022.

SURLES, J.; PADGETT, W. Inference for reliability and stress-strength for a scaled Burr type X distribution. *Lifetime data analysis*, Springer, v. 7, p. 187–200, 2001.

TAHIR, M. H.; HUSSAIN, M. A.; CORDEIRO, G. M. A new flexible generalized family for constructing many families of distributions. *Journal of Applied Statistics*, v. 49, n. 7, p. 1615–1635, 2022.

TARVIRDIZADE, B. The Lomax-Lindley distribution: Properties and applications to lifetime data. *Communications Faculty of Sciences University of Ankara Series A1 Mathematics and Statistics*, Ankara University, v. 70, n. 2, p. 965–983, 2021.

THOMAS, M.; ROOTZÉN, H. Real-time prediction of severe influenza epidemics using extreme value statistics. *Journal of the Royal Statistical Society Series C: Applied Statistics*, v. 71, n. 2, p. 376–394, 03 2022.

TIAN, N.; ZHENG, J.-X.; GUO, Z.-Y.; LI, L.-H.; XIA, S.; LV, S.; ZHOU, X.-N. Dengue incidence trends and its burden in major endemic regions from 1990 to 2019. *Tropical Medicine and Infectious Disease*, MDPI, v. 7, n. 8, p. 180, 2022.

TIPIRNENI, R.; SCHMIDT, H.; LANTZ, P. M.; KARMAKAR, M. Associations of 4 geographic social vulnerability indices with US COVID-19 incidence and mortality. *American Journal of Public Health*, American Public Health Association, v. 112, n. 11, p. 1584–1588, 2022.

TRAN, Q. V.; KUKAL, J. A novel heavy tail distribution of logarithmic returns of cryptocurrencies. *Finance Research Letters*, Elsevier, v. 47, p. 102574, 2022.

VASCONCELOS, J. C. S.; CORDEIRO, G. M.; ORTEGA, E. M.; ARAÚJO, E. G. The new odd log-logistic generalized inverse Gaussian regression model. *Journal of Probability and Statistics*, Hindawi Limited, v. 2019, p. 1–13, 2019.

VASCONCELOS, J. C. S.; CORDEIRO, G. M.; ORTEGA, E. M.; SAULO, H. Parametric and partially linear regressions for agricultural economy data. *Communications in Statistics-Theory and Methods*, Taylor & Francis, p. 1–25, 2022.

VASCONCELOS, J. C. S.; CORDEIRO, G. M.; ORTEGA, E. M. M.; REZENDE, É. M. d. A new regression model for bimodal data and applications in agriculture. *Journal of Applied Statistics*, Taylor & Francis, v. 48, n. 2, p. 349–372, 2021.

VASCONCELOS, J. C. S.; CORDEIRO, G. M.; ORTEGA, E. M. M.; BIAGGIONI, M. A. M. The parametric and additive partial linear regressions based on the generalized odd log-logistic log-normal distribution. *Communications in Statistics - Theory and Methods*, Taylor Francis, v. 51, n. 11, p. 3480–3507, 2022.

VIGAS, V. P.; ORTEGA, E. M.; SUZUKI, A. K.; CORDEIRO, G. M.; JUNIOR, P. C. dos S. The generalized odd log-logistic-G regression with interval-censored survival data. *Journal of Applied Statistics*, Taylor & Francis, p. 1–22, 2023.

VILA, R.; ÇANKAYA, M. N. A bimodal Weibull distribution: Properties and inference. *Journal of Applied Statistics*, Taylor & Francis, v. 49, n. 12, p. 3044–3062, 2022.

WANG, P.; PAL, S. A two-way flexible generalized gamma transformation cure rate model. *Statistics in Medicine*, Wiley Online Library, v. 41, n. 13, p. 2427–2447, 2022.

YANG, Y.; JAVANROODI, K.; NIK, V. M. Climate change and renewable energy generation in Europe - long-term impact assessment on solar and wind energy using high-resolution future climate data and considering climate uncertainties. *Energies*, MDPI, v. 15, n. 1, p. 302, 2022.

YELIN, I.; KATZ, R.; HERZEL, E.; BERMAN-ZILBERSTEIN, T.; BEN-TOV, A.; KUINT, J.; GAZIT, S.; PATALON, T.; CHODICK, G.; KISHONY, R. Associations of the BNT162b2 COVID-19 vaccine effectiveness with patient age and comorbidities. *medrxiv*, Cold Spring Harbor Laboratory Press, p. 2021–03, 2021.

ZEGHDOUDI, H.; NEDJAR, S. Gamma Lindley distribution and its application. *Journal of Applied Probability and Statistics*, v. 11, n. 1, p. 129–138, 2016.