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

The parametric and semiparametric regression models based on the generalized odd log-logistic family

Fábio Prataviera

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

Piracicaba
2020

Fábio Prataviera
Degree in Statistics

The parametric and semiparametric regression models based on the generalized odd log-logistic family

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

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

Piracicaba
2020

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

Prataviera, Fábio
The parametric and semiparametric regression models based on the generalized odd log-logistic family / Fábio Prataviera. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2020 .
130 p.

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

1. Dados censurados 2. Fração de cura 3. Inflação de zeros 4. Spline cúbico 5. Simulação . I. Título.

DEDICATORATION

To my parents,
João Batista Prataviera and **Zilda Aparecida Tolon Prataviera**, for all the love, patience and dedication they have for me.

To my brother,
Marcelo Henrique Prataviera for the friendship, laughter, patience and trust in me.

To my uncle,
Gilberto Aparecido Prataviera for the friendship, for the teachings and for all the support dedicated to me.

To them,
I dedicate this work.

ACKNOWLEDGMENTS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Ao meu orientador, Prof. Dr. Edwin Moises Marcos Ortega, pelo entusiasmo, ajuda, reconhecimento, apoio e dedicação para o desenvolvimento deste trabalho.

Aos professores da minha banca de qualificação, Profª. Drª. Renata Alcarde, Profª. Drª. Elizabethe Mie Hashimoto e Prof. Dr. Vicente Garibay Cancho, pelas sugestões que contribuiram muito para o aperfeiçoamento deste trabalho.

Ao Prof. Dr. Gauss Cordeiro, pela solicitude e por toda contribuição no desenvolvimento deste trabalho.

Ao doutourando Antonio Marcos Miranda Silva e a Profª. Drª Elke Jurandy Bran Nogueira Cardoso, pela parceria e contribuição no Capítulo 4 deste trabalho.

Ao Prof. Dr. Idemauro Antonio Rodrigues de Lara, pela supervisão no Programa de Aperfeiçoamento de Ensino - PAE.

À todos os professores do curso de Pós-graduação em Estatística e Experimentação Agronômica, pelos ensinamentos e possibilidades que contribuíram para minha formação acadêmica.

Às secretárias do Departamento de Ciências Exatas, Luciane Brajão e Solange de Assis Paes Sabadin, e aos técnicos de informática, Eduardo Bonilha e Jorge Alexandre Wiendl, pela ajuda e boa vontade.

À minha namorada, Aline Martineli Batista, por todo carinho, apoio e compreensão, tornando assim os meus dias mais leves e felizes. Obrigado.

À todos os amigos dos cursos de mestrado e doutorado do Programa de Pós-graduação em Estatística e Experimentação Agronômica, pelos momentos de estudos, descontração, a atenção e amizade.

Enfim, a todos que me ajudaram de forma direta ou indireta para o desenvolvimento deste trabalho.

CONTENTS

RESUMO

Modelos de regressão parametricos e semiparametricos baseados na família generalizada odd log-logística

Nesse trabalho foram realizadas diferentes análises via modelos de regressão considerando a família geradora de novas distribuições, denominada de *generalizada odd log-logística-G* (GOLL-G). As distribuições nesta família apresentam maior flexibilidade, como por exemplo, funções de densidades bimodais. Com base na família GOLL-G, foram propostos: modelos de regressão com diferentes estruturas de regressão; modelo semi-paramétrico inflacionado de zeros modelando os parâmetros via splines penalizados; Para todas as abordagens o recurso computacional para implementação dos modelos foi o *software* R, sendo apresentados trechos de comandos ao longo do documento assim como breve descrições dos códigos usados. Os resultados obtidos nas aplicações mostram que o modelo proposto pode ser uma alternativa interessante, principalmente quando os dados apresentam assimetria e bimodalidade.

Palavras-chave: Dados censurados, Fração de cura, Inflação de zeros, Spline cúbico, Simulação

ABSTRACT

The parametric and semiparametric regression models based on the generalized odd log-logistic family

In this work, several analyzes were performed through regression models considering the family of new distributions, called *generalized odd log-logistic-G* (GOLL-G), the distributions in this family have greater flexibility, such as functions of bimodal densities. Based on the GOLL-G family, we proposed: regression models with different regression structures; inflated semi-parametric model of zeros modeling of the parameters via penalized splines; For all the modeling approaches presented, the computational resource for the implementation of the models was *software* R, throughout the document as well as brief descriptions of the codes used. The results obtained in the applications show that the proposed model can be an interesting alternative, especially when the data present asymmetry and bimodality.

Keywords: Censored data, Cure rate, Cubic spline, Zero inflated, Simulation

1 INTRODUCTION

Regression analysis is a commonly used statistical technique applied in many scientific fields. The linear regression model with normal distribution is generally used to model data having symmetric distribution. However, various phenomena cannot always be modeled with the normal distribution, be it for the lack of symmetry, the existence of bimodality or the presence of atypical values.

In past decades, when the phenomenon of interest did not satisfy the assumption of normality of the response variable, some type of transformation was applied at least to obtain symmetric behavior of the data. However, recently it has become more attractive to propose new regression models to model different types of data.

In this work we use regression model to solve problems in different areas. For example, in survival analysis the study of the lifetime of patients with a particular disease and the study of the failure time of an electronic component. The study of times is called survival analysis in the medical area and reliability analysis in the industrial area.

There are also situations where continuous data can include a high percentage of zeros. In these situations, continuous distributions can not be used. The data that contain excessive zeros can be analyzed by a mixture of two distributions: a continuous distribution (with positive support) and a degenerate distribution at zero, i.e. a model whose mixed discrete-continuous probability and distribution functions.

Among the different proposed models and families of distribution, it is notable that only a small number take bimodal forms. In this work a new model based on the *generalized odd log-logistic - G* (Cordeiro *et al.* 2017) family is proposed. We consider the bade distribution the Maxwell distribution.

The Maxwell (or Maxwell-Boltzmann) distribution is an important model in physics, chemistry and statistical mechanics. It forms the basis of the kinetic energy of gases and explains several fundamental properties of gases including pressure and diffusion. In statistical mechanics, it is related to properties of molecules in thermal equilibrium from the microscopic perspective. The Maxwell distribution is also important in kinetic translational energies for molecules. For example, Prigogine and Xhrouet (1949) discussed this distribution for chemical reactions in gases and Brilliantov and Poschel (2000) studied its deviations in granular gases with constant coefficient of restitution.

In recent years, the Maxwell distribution has been used to model failure times in survival and reliability analysis and some of its extended forms have been investigated. Krishna *et al.* (2012) addressed reliability estimation in the Maxwell distribution with progressively type-II censored data, Kazmi *et al.* (2012) explored a heterogeneous population by means of two mixture components of Maxwell distributions, Tomer and Panwar *et al.* (2015) considered point and interval estimation procedures for the Maxwell distribution in the presence of type-I progressively hybrid censored data, Dey *et al.* (2016) presented its structural properties and different methods of estimation, Iriarte *et al.* (2016) defined the gamma-Maxwell distribution and, more recently, Venegas *et al.* (2017) proposed the transmuted exponentiated Maxwell distribution. However, none of these papers deal with bimodality to real data and do not even present regression models for the extensions of the Maxwell distribution. We aim to fill up this gap.

Based on the proposed model, we try to solve problems from different areas based on regression models. However, in many situations the relationship between the response variable and the explanatory (or covariate) variable has no linear relationship. This can often make it difficult to explain this relationship. In this work we propose the generalized odd log-logistic Maxwell parametric and semiparametric regression models to solve the problems above. In addition, to solve the issue of nonlinear behavior and in

order to obtain a more flexible model for the data we use cubic splines in this work. Thus, the inclusion of cubic splines in the model requires, in addition to descriptive and exploratory analyzes, diagnostic analyzes to assess the suitability of the model.

Another situation that occurs in many studies in several fields aim to determine how a set of explanatory variables influence other variables expressed as ratios or proportions, i.e., random experiments that produce results in the interval $(0, 1)$. Several researchers tried to model this type of data. For example, Ferrari and Cribari-Neto (2004) pioneered a regression in which the parameters are interpreted as mean and precision, Bayes *et al.* (2012) proposed a robust regression for proportions based on the beta rectangular distribution, Lemonte and Bazán (2016) defined a class of Johnson SB distributions and its associated regression for rates and proportions, Mazucheli *et al.* (2019) proposed a unit-Lindley distribution and its associated regression for proportional data. In these terms, our main aim is to propose a regression based on the generalized odd log-logistic beta ("GOLLBE" for short) distribution to model proportional data with bimodality.

All computational scripts of the new regression model were implemented in the `R` software using the `gamlss` package (Stasinopoulos and Rigby, 2007).

1.1 Objetive

The objetive of this work is to propose a new probability distribution based on the family proposed by Cordeiro *et al.* (2017). In this way, study of the properties and characteristics of such models are studied. In addition, to present applications of the proposed models in different types of data considering a regression structure.

1.2 Work Organization

This thesis is organized as follows. In Chapter 2, a new probability distribution called the generalized odd log-logistic Maxwell, whose main advantage related to other competitive distribution is modeling bimodal, asymmetric and heavy tails data. Some properties of this model are presented. Three applications to real data sets in engineering are presented. The Chapter 3, considers the generalized odd log-logistic Maxwell distribution with a semiparametric regression structure applied to censored and uncensored data. The non-parametric term of the model studied is approximated considering the cubic spline functions. In Chapter 4, considers the new zero adjusted generalized odd log-logistic Maxwell distribution with a semiparametric regression structure applied using data from an experiment conducted to assess the soil microbiology in a sugarcane field. In Chapter 4, considers the new zero adjusted generalized odd log-logistic Maxwell distribution with a semiparametric regression structure applied using data from an experiment conducted to assess the soil microbiology in a sugarcane field. The generalized odd log-logistic Maxwell rate cure semiparametric regression models is presented in Chapter 5. An application in prostate cancer data is performed, considering a regression structure joint in time and the cure rate. In Chapter 6, a reparametrization in the median of the generalized odd log-logistic Maxwell model is proposed with application in agricultural data. In Chapter 7, is proposed a new continous distribution in the interval (0,1). Some properties of this model are presented. For the regression model, studies of diagnotics and residual analysis are performed. Finally, some considerations and perspectives for future work are presented in Chapter 8.

References

Bayes, C.L., Bazán, J.L. and García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis*, **4**, 841-866.

Brilliantov, N.V. and Poschel, T. (2000). *In Granular Gases*. Springer, Berlin.

Cordeiro, G.M., Alizadeh, M., Ozel, G., Hosseini, B., Ortega, E.M.M. and Altun, E. (2017). The generalized odd log-logistic family of distributions: properties, regression models and applications. *Journal of Statistical Computation and Simulation*, **87**, 908-932.

Dey, S., Dey, T., Ali, S. and Mulekar, M.S. (2016). Two-parameter Maxwell distribution: Properties and different methods of estimation. *Journal of Statistical Theory and Practice*, **10**, 291-310.

Ferrari, S.L.P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799-815.

Iriarte, Y.A., Astorga, J.M., Bolfarine, H. and Gómez, H.W. (2017). Gamma-Maxwell distribution. *Communications in Statistics-Theory and Methods*, **46**, 4264-4274.

Kazmi, S., Aslam, M. and Ali, S. (2012). On the Bayesian estimation for two component mixture of Maxwell distribution, assuming type I censored data. *SOURCE International Journal of Applied Science and Technology*, **2**, 197-218.

Krishna, H. and Malik, M. (2012). Reliability estimation in Maxwell distribution with progressively type-II censored data. *Journal of Statistical Computation and Simulation*, **82**, 623-641.

Lemonte, A.J. and Bazán, J.L. (2016). New class of Johnson distributions and its associated regression model for rates and proportions. *Biometrical Journal*, **58**, 727-746.

Mazucheli, J., Menezes, A.F.B. and Chakraborty, S. (2019). On the one parameter unit-Lindley distribution and its associated regression model for proportion data. *Journal of Applied Statistics*, **46**, 700-714.

Prigogine, I., Xhrouet, E. (1949). On the perturbation of Maxwell distribution function by chemical reactions in gases. *Physica*, **15**, 913-932.

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

Tomer, S.K. and Panwar, M. S. (2015). Estimation procedures for Maxwell distribution under type-I progressive hybrid censoring scheme. *Journal of Statistical Computation and Simulation*, **85**, 339-356.

Venegas, O., Iriarte, Y.A., Astorga, J.M., Borger, A., Bolfarine, H. and Gómez, H.W. (2017). A New Generalization of the Maxwell Distribution. *Applied Mathematics and Information Sciences*, **11**, 867-876.

2 THE GENERALIZED ODD LOG-LOGISTIC MAXWELL DISTRIBUTION WITH APPLICATIONS IN ENGINEERING

Abstract: We define the generalized odd log-logistic Maxwell distribution and propose a parametric regression model based on the new distribution with three systematic components for its parameters. Some properties and maximum likelihood estimation are addressed and various simulations for different parameter settings, systematic components and sample sizes are performed. We present a simulation study to verify the adequacy of the normal approximation to the quantile residuals in the regression model. Three applications to real data sets in engineering (strength and mechanics of materials) prove empirically the usefulness of the proposed models.
Keywords: Maximum likelihood; Maxwell distribution; Odd log-logistic distribution; Regression model; Simulation.

2.1 Introduction

The Maxwell (or Maxwell-Boltzmann) distribution is an important model in physics, chemistry and statistical mechanics. It forms the basis of the kinetic energy of gases and explains several fundamental properties of gases including pressure and diffusion. In statistical mechanics, it is related to properties of molecules in thermal equilibrium from the microscopic perspective. The Maxwell distribution is also important in kinetic translational energies for molecules. For example, Prigogine and Xhrouet (1949) discussed this distribution for chemical reactions in gases and Brilliantov and Poschel (2000) studied its deviations in granular gases with constant coefficient of restitution.

In recent years, the Maxwell distribution has been used to model failure times in survival and reliability analysis and some of its extended forms have been investigated. Krishna *et al.* (2012) addressed reliability estimation in the Maxwell distribution with progressively type-II censored data and Kazmi *et al.* (2012) explored a heterogeneous population by means of two mixture components of Maxwell distributions. Tomer and Panwar *et al.* (2015) considered point and interval estimation procedures for the Maxwell distribution in the presence of type-I progressively hybrid censored data, Dey *et al.* (2016) presented its structural properties and different methods of estimation, Iriarte *et al.* (2016) defined the gamma-Maxwell distribution and, more recently, Venegas *et al.* (2017) proposed the transmuted exponentiated Maxwell distribution. However, none of these papers deal with bimodality to real data and do not even present regression models for the extensions of the Maxwell distribution. We aim to fill up this gap.

Many scientific studies involve data with bimodal characteristics of continuous random variables, in which the usual choice is to employ a mixture of distributions. However, the most common mixtures of distributions have a large number of parameters, thus making the estimation of these parameters complicated. For example, engineers study materials to learn their properties and the problems that the materials can cause. In this respect, ceramic materials are generally composed of a combination of metallic and non-metallic elements (forming oxides, nitrides and carbides), and are more resistant to high temperatures and severe environments than metals and polymers. In the applications, we study chemical compounds with four levels (ZrO_2, $ZrO_2 - TiB_2$, Si_3N_4 and glass). The behavior of each level is shown in Figure 2.1, where the asymmetry and bimodality of the data can be noted.

In this context, we define a new distribution called the *generalized odd log-logistic Maxwell* ("GOLLMax" for short), whose main advantage related to other competitive distribution is modeling

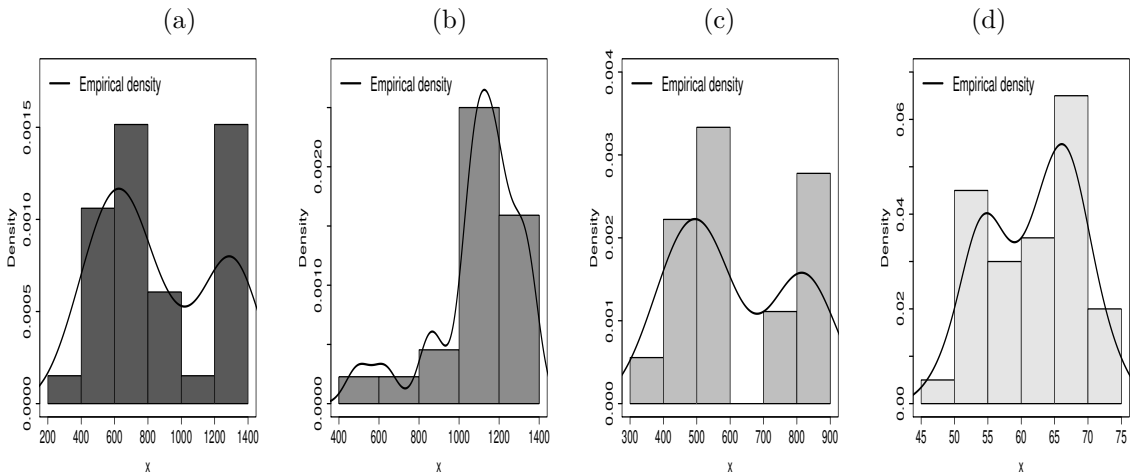

Figure 2.1. Histograms and empirical densities. (a) ZrO_2. (b) $ZrO_2 - TiB_2$. (c) Si_3N_4. (d) Glass.

bimodal, asymmetric and heavy tails data. It includes as special cases the Maxwell, exponentiated Maxwell (EMax) and odd log-logistic Maxwell (OLLMax) distributions, among others. We derive some mathematical properties of the proposed distribution. In practice, it is quite common situations where there are some explanatory variables associated with the response random variable. For example, in industry, the failure time of an equipment can be influenced by the voltage level to which the equipment is subjected. In the medical field, a patient's survival time can be related to the type of tumor and the amount of hemoglobin in the blood. In general, we study the effects of these explanatory variables on the response variable by means of a regression model.

In the second part of this chapter, we propose a regression model based on the new distribution and present some global influence measures. In addition, we develop residual analysis based on quantile residuals. For different parameter settings and sample sizes, various simulation studies are performed and the empirical distribution of these residuals is compared with the standard normal distribution. The simulation results indicate that the empirical distribution of the quantile residuals is in agreement with the standard normal distribution.

The rest of the chapter is organized as follows. In Section 2.2, we introduce the new distribution and subsection 2.2.1 some mathematical properties of the GOLLMax distribution are presented. We propose the GOLLMax regression model with three systematic structures in Section 2.3 and evaluate the performance of the maximum likelihood estimators (MLEs) of the model parameters by means of a simulation study. In Section 2.5, we investigate the case-deletion diagnostic measure and define quantile residuals for the fitted model. Further, we perform various simulations for these residuals. In Section 2.6, we provide three applications to real data to illustrate the flexibility of the new models. Finally, some concluding remarks are offered in Section 2.7.

2.2 The model definition

The cumulative distribution function (cdf) of the Maxwell random variable W is given by

$$G(x;\mu) = \gamma_1\left(\frac{3}{2}, \frac{x^2}{\mu^2}\right), \; x > 0, \tag{2.1}$$

where x denotes the molecule speed, $\mu > 0$ is a scale parameter depending on three quantities (Boltzmann constant, temperature and mass of a molecule), $\gamma_1(p, y) = \gamma(p, y)/\Gamma(p)$ is the incomplete gamma function ratio, $\gamma(p, y) = \int_0^u z^{p-1}\mathrm{e}^{-z}\mathrm{dz}$ is the incomplete gamma function and $\Gamma(\cdot)$ is the gamma function.

The probability density function (pdf) of W is given by

$$g(x;\mu) = \frac{4}{\sqrt{\pi}}\,\frac{x^2}{\mu^3}\,\exp\left(-\frac{x^2}{\mu^2}\right), \quad x > 0. \tag{2.2}$$

The expectation and the variance of W are $E(W) = 2\mu/\sqrt{\pi}$ and $Var(W) = (3\pi - 8)\mu^2/(2\pi)$. We have the following relation of the μ parameter, $\mu = 1/\sqrt{\frac{m}{2K_BT}}$, where K_B is the Boltzmann constant, T is temperature and m is the molecular mass. Thus, considering such quantities, the probability density function (2.2) is given by

$$g(x) = \frac{4}{\sqrt{\pi}}\,\frac{m}{2K_BT}x^2\,\exp\left(-\frac{m\,x^2}{2K_BT}\right),$$

with average velocity of the molecules at a certain temperature is $\sqrt{\frac{8K_BT}{m\pi}}$.

The idea of the GOLLMax distribution follows the generator. Let $G(x;\boldsymbol{\gamma})$ be a baseline cdf having a $p \times 1$ vector $\boldsymbol{\gamma}$ of unknown parameters. Cordeiro *et al.* (2017) defined the cdf of a wider generator called the *generalized odd log-logistic-G* ("GOLL-G") family, by integrating the log-logistic density function, namely

$$F(x;\sigma,\nu,\boldsymbol{\gamma}) = \int_0^{\frac{G(x;\boldsymbol{\gamma})^\sigma}{1-G(x;\boldsymbol{\gamma})^\sigma}} \frac{\nu\,w^{\nu-1}}{(1+w^\nu)^2}dw = \frac{G(x;\boldsymbol{\gamma})^{\sigma\nu}}{G(x;\boldsymbol{\gamma})^{\sigma\nu} + [1-G(x;\boldsymbol{\gamma})^\sigma]^\nu}, \tag{2.3}$$

where $\sigma > 0$ and $\nu > 0$ are two extra shape parameters. Equation (2.3) includes as special cases the *odd log-logistic-G* (OLL-G) family introduced by Gleaton and Lynch (2006) and the exponentiated-G (exp-G) class when $\sigma = 1$ and $\nu = 1$, respectively. Further, the G distribution is the basic exemplar when $\sigma = \nu = 1$.

We define the cdf of the three-parameter GOLLMax distribution by inserting (2.1) in equation (2.3), give by

$$F(x;\mu,\sigma,\nu) = \frac{\gamma_1^{\sigma\nu}(3/2, x^2/\mu^2)}{\gamma_1^{\sigma\nu}(3/2, x^2/\mu^2) + [1-\gamma_1^\sigma(3/2, x^2/\mu^2)]^\nu}. \tag{2.4}$$

Henceforth, if X is a random variable with cdf (2.4), we write $X \sim \mathrm{GOLLMax}(\mu,\sigma,\nu)$.

By differentiating (2.3) and inserting (2.1) and (2.2), we obtain the density function of X (for $x > 0$) as

$$f(x;\mu,\sigma,\nu) = \frac{4\sigma\nu}{\sqrt{\pi}\mu^3}\,x^2\,\exp\left(-\frac{x^2}{\mu^2}\right)\frac{\gamma_1^{\sigma\nu-1}(3/2, x^2/\mu^2)\left[1-\gamma_1^\sigma(3/2, x^2/\mu^2)\right]^{\nu-1}}{\left\{\gamma_1^{\sigma\nu}(3/2, x^2/\mu^2) + [1-\gamma_1^\sigma(3/2, x^2/\mu^2)]^\nu\right\}^2}. \tag{2.5}$$

The hazard rate function (hrf) of X is $h(x) = f(x)/[1 - F(x)]$. The GOLLMax model contains as special cases the following distributions:

- For $\sigma = 1$, it gives the (new) OLLMax distribution.
- For $\nu = 1$, it yields the (new) EMax distribution.
- The Maxwell distribution is as a basic exemplar when $\sigma = \nu = 1$.

Some plots of the density and hrf of X for selected parameter values, including well-known distributions, are displayed in Figures 2.2 and 2.3, respectively. A characteristic of the proposed distribution is that its hrf can be bathtub shaped, monotonically (increasing or decreasing), unimodal, increasing-decreasing-increasing shaped, among others, depending basically on the parameter values.

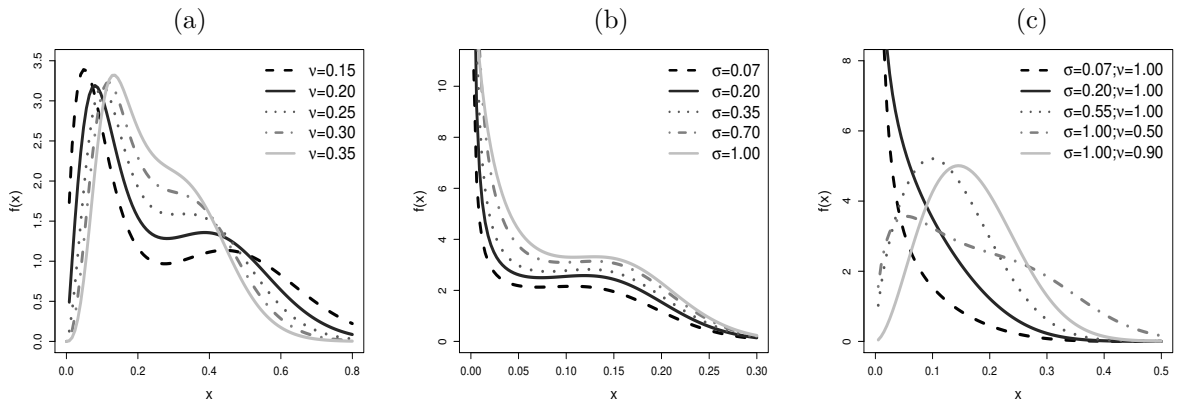

Figure 2.2. Plots of the GOLLMax density for some parameter values. (a) For different values of ν with $\sigma = 3.45$ and $\mu = 0.15$. (b) For different values of σ with $\nu = 0.30$ and $\mu = 0.07$. (c) For different values of ν and σ with $\mu = 0.15$.

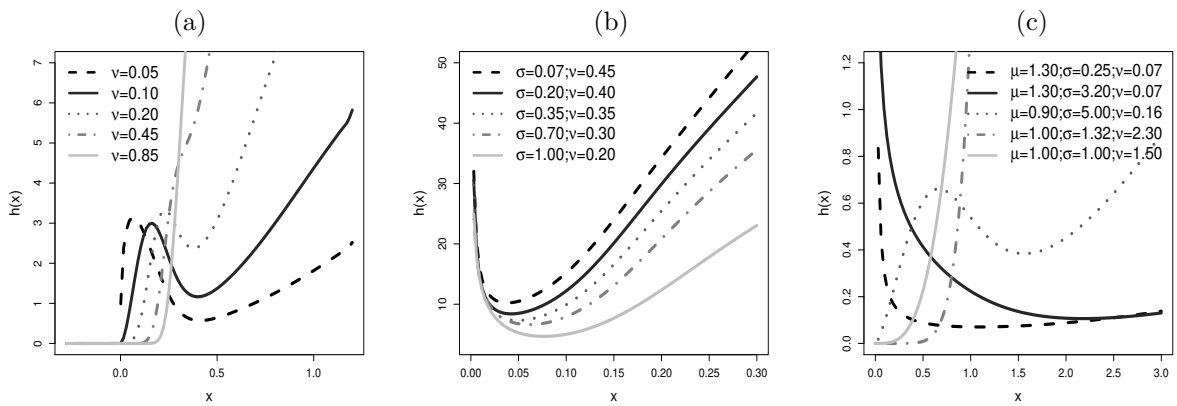

Figure 2.3. Plots of the GOLLMax hrf for some parameter values. (a) For different values of ν with $\sigma = 9.00$ and $\mu = 0.20$. (b) For different values of ν and σ with $\mu = 0.07$. (c) For different values of μ σ and ν.

Equation (2.4) has tractable properties especially for simulations, since its quantile function (qf) takes the simple form

$$x = Q_{\text{Max}}\left(\left[\frac{\left(\frac{u}{1-u}\right)^{\frac{1}{\nu}}}{1+\left(\frac{u}{1-u}\right)^{\frac{1}{\nu}}}\right]^{\frac{1}{\sigma}}\right), \tag{2.6}$$

where $Q_{\text{Max}}(u) = G^{-1}(\mu; u)$ is the qf of the Maxwell distribution.

Further, we can write

$$x = Q_{\text{Max}}(w; \mu) = \mu\sqrt{\gamma_1^{-1}(3/2, w)},$$

where $w = \left[\frac{u^{\frac{1}{\nu}}}{(1-u)^{\frac{1}{\nu}}+u^{\frac{1}{\nu}}}\right]^{\frac{1}{\sigma}}$ and $\gamma_1^{-1}(3/2, w)$ is the inverse of the upper gamma regularized function. For more details, see `http://functions.wolfram.com/GammaBetaErf/InverseGammaRegularized/`.

2.2.1 Structural properties

We derive here some mathematical properties of $X \sim \text{GOLLMax}(\mu, \sigma, \nu)$. First, we introduce some notation. Let $\Pi(x; \mu, k, \beta) = \gamma_1\left(k, [\frac{x}{\mu}]^\beta\right)$ (for $x > 0$) be the cdf of the *generalized gamma* (GG) (Stacy, 1962) distribution with shape parameters $k > 0$ and $\beta > 0$ and scale parameter $\mu > 0$. The corresponding density function, say $\pi(x; \mu, k, \beta)$, is

$$\pi(x, \mu, k, \beta) = \frac{\beta}{\mu\,\Gamma(k)} \left(\frac{x}{\mu}\right)^{k\beta-1} \exp\left[-\left(\frac{x}{\mu}\right)^\beta\right]. \tag{2.7}$$

Clearly, the Maxwell density (2.2) follows from (2.7) as $g(x; \mu) = \pi(x; \mu, 3/2, 2)$.

Theorem 1: We can express the density $f(x; \mu, \sigma, \nu)$ of X as a linear combination of GG densities

$$f(x; \mu, \sigma, \nu) = \sum_{m,i=0}^{\infty} p_{m,i}\ \pi(x; \mu, k^*, 2), \tag{2.8}$$

where the GG densities have common parameters μ and 2 and varying shape parameter $k^* = k^*(m, i) = 3(m + 1)/2 + i$, and the coefficients $p_{m,i} = p_{m,i}(\mu, \sigma, \nu)$ are functions of the model quantities (defined below) given by

$$p_{m,i} = \frac{2(m+1)\,\mu^{3m+2i}\ w_{m+1}\,c_{m,i}}{\sqrt{\pi}}\,\Gamma\left(\frac{3m}{2} + i\right).$$

Proof Theorem 1:

For any real non-integer $\lambda > 0$, the generalized binomial theorem

$$\gamma_1(3/2, x^2/\mu^2)^\lambda = \left[1 - \left\{1 - \gamma_1(3/2, x^2/\mu^2)\right\}\right]^\lambda = \sum_{j=0}^{\infty} (-1)^j \binom{\lambda}{j} \left\{1 - \gamma_1(3/2, x^2/\mu^2)\right\}^j,$$

holds, and it is always convergent since $0 < \gamma_1(3/2, x^2/\mu^2) < 1$. Hence,

$$\gamma_1(3/2, x^2/\mu^2)^\lambda = \sum_{j=0}^{\infty} \sum_{m=0}^{j} (-1)^{j+m} \binom{\lambda}{j} \binom{j}{m} \gamma_1(3/2, x^2/\mu^2)^m.$$

By substituting $\sum_{j=0}^{\infty} \sum_{m=0}^{j}$ for $\sum_{m=0}^{\infty} \sum_{j=m}^{\infty}$ and after some algebra, we obtain

$$\gamma_1(3/2, x^2/\mu^2)^\lambda = \sum_{m=0}^{\infty} s_m(\lambda)\,\gamma_1(3/2, x^2/\mu^2)^m, \tag{2.9}$$

where (for $m \geq 0$)

$$s_m(\lambda) = \sum_{j=m}^{\infty} (-1)^{r+m} \binom{\lambda}{j} \binom{j}{m}.$$

By using (2.9), the numerator of (2.4) can be expanded as

$$\gamma_1^{\sigma\nu}(3/2, x^2/\mu^2) = \sum_{m=0}^{\infty} s_m(\sigma\nu)\,\gamma_1(3/2, x^2/\mu^2)^m \tag{2.10}$$

where $s_m(\sigma\nu)$ comes from a previous quantity. By using the generalized binomial theorem and (2.9), one part of the denominator (2.4) can be written as

$$\begin{aligned} \left[1 - \gamma_1^\sigma(3/2, x^2/\mu^2)\right]^\nu &= 1 + \sum_{j=1}^{\infty} (-1)^j \binom{\nu}{j} \gamma_1^{j\sigma}(3/2, x^2/\mu^2) \\ &= 1 + \sum_{m=0}^{\infty} t_m(\sigma, \nu)\,\gamma_1(3/2, x^2/\mu^2)^m, \end{aligned}$$

where $t_m(\sigma,\nu) = \sum_{j=1}^{\infty}(-1)^j \binom{\nu}{j} s_m(j\sigma)$.

The denominator of (2.4) can be determined from (2.10) and the last power series as

$$\gamma_1^{\sigma\nu}(3/2, x^2/\mu^2) + \left[1 - \gamma_1^{\sigma}(3/2, x^2/\mu^2)\right]^{\nu} = \sum_{m=0}^{\infty} v_m(\sigma\nu)\, \gamma_1(3/2, x^2/\mu^2)^m, \tag{2.11}$$

where $v_0(\sigma,\nu) = 1 + s_0(\sigma\nu) + t_0(\sigma,\nu)$ and $v_m(\sigma,\nu) = s_m(\sigma\nu) + t_m(\sigma,\nu)$ for $m \geq 1$.

Combining (2.10) and (2.11), we can express (2.4) as

$$F(x;\mu,\sigma,\nu) = \frac{\sum_{m=0}^{\infty} s_m(\sigma\nu)\, \gamma_1(3/2, x^2/\mu^2)^m}{\sum_{m=0}^{\infty} v_m(\sigma,\nu)\, \gamma_1(3/2, x^2/\mu^2)^m}.$$

The ratio of the two power series in the last equation reduces to

$$F(x;\mu,\sigma,\nu) = \sum_{m=0}^{\infty} w_m\, \gamma_1(3/2, x^2/\mu^2)^m, \tag{2.12}$$

where the coefficients w_m's (for $m \geq 1$) are determined from the recurrence equation

$$w_m = w_m(\sigma,\nu) = v_0(\sigma,\nu)^{-1} \left[s_m(\sigma\nu) - \sum_{r=1}^{m} v_r(\sigma,\nu)\, w_{m-r}(\sigma,\nu) \right]$$

and $w_0 = w_0(\sigma,\nu) = s_0(\sigma\nu)/v_0(\sigma,\nu)$.

By differentiating (2.12), we can rewrite (2.5) as

$$f(x;\mu,\sigma,\nu) = \sum_{m=0}^{\infty} w_{m+1}\, h_{m+1}(x;\mu), \tag{2.13}$$

where $h_{m+1}(x;\mu) = (m+1) g(x;\mu)\, \gamma_1(3/2, x^2/\mu^2)^m$ is the EMax density with power parameter $(m+1)$ (for $m \geq 0$). Then, we have

$$h_{m+1}(x;\mu) = \frac{4\,(m+1)}{\sqrt{\pi}\,\mu^3}\; x^2\, \exp\left(-\frac{x^2}{\mu^2}\right)\, \gamma_1(3/2, x^2/\mu^2)^m.$$

Further, we adopt the power series for the incomplete gamma function ratio

$$\gamma_1(3/2, x^2/\mu^2) = \sum_{i=0}^{\infty} a_i\, x^{2i+3}, \tag{2.14}$$

where $a_i = a_i(\mu) = \frac{(-1)^i}{(3/2+i)\,\mu^{2i+3}\,\Gamma(3/2)\,i!}$ (for $i \geq 0$).

By application of an equation in Section 0.314 of Gradshteyn and Ryzhik (2000) for power series raised to powers, we can write (for any m positive integer)

$$\left(\sum_{i=0}^{\infty} a_i\, z^i\right)^m = \sum_{i=0}^{\infty} c_{m,i}\, z^i, \tag{2.15}$$

where the coefficients $c_{m,i} = c_{m,i}(\mu)$ are determined by $c_{m,0} = a_0^m$ and, for $i = 1, 2, \ldots$, from the recurrence relation

$$c_{m,i} = (i\, a_0)^{-1} \sum_{r=1}^{i} [(m+1)r - i]\, a_r\, c_{m,i-r}.$$

So, the coefficient $c_{m,i}$ can follow from $c_{m,0}, \cdots, c_{m,i-1}$ and then from $a_0, \cdots, a_i$ and

Further, we have from equation (2.15)

$$\gamma_1(3/2, x^2/\mu^2)^m = \sum_{i=0}^{\infty} c_{m,i}\, x^{3m+2i}. \tag{2.16}$$

Combining (2.9) and (2.16), the EMax density can be expanded as

$$h_{m+1}(x;\mu) = \frac{4}{\sqrt{\pi}\,\mu^3} \sum_{i=0}^{\infty}(m+1)\,c_{m,i}\,x^{3m+2i+2}\,\exp\left(-\frac{x^2}{\mu^2}\right).$$

By inserting this expression in Equation (2.13) gives

$$f(x;\mu,\sigma,\nu) = \frac{4}{\sqrt{\pi}\,\mu^3} \sum_{m,i=0}^{\infty}(m+1)\,w_{m+1}\,c_{m,i}\,x^{3m+2i+2}\,\exp\left(-\frac{x^2}{\mu^2}\right).$$

Finally, we can rewrite $f(x;\mu,\sigma,\nu)$ as a linear combination of GG densities with two common parameters 2 and μ and the third parameter $k^* = k^*(m,i) = 3(m+1)/2+i$ as in (2.8), and the coefficients $p_{m,i} = p_{m,i}(\mu,\sigma,\nu)$ are functions of previous quantities given by

$$p_{m,i} = \frac{2(m+1)\,\mu^{3m+2i}}{\sqrt{\pi}}\,\Gamma\left(3m/2+i\right)\,w_{m+1}\,c_{m,i}.$$

The linear representation (2.8) becomes very useful in deriving some mathematical properties for the GOLLMax distribution using well-known GG properties. We can adopt at most ten terms in (2.8) to provide accurate results for most analytical platforms.

Corollary 1 The nth moment of X takes the form

$$\mu'_n = E(X^n) = \mu^n \sum_{m,i=0}^{\infty} p_{m,i}\,\frac{\Gamma\left(3[m+1]/2+i+n/2\right)}{\Gamma\left(3[m+1]/2+i\right)}. \tag{2.17}$$

Proof Corollary 1:

The nth ordinary moment of the GG pdf $\pi(x;\mu,k,\beta)$ is known to be $\delta'_{n,GG} = \mu^n\,\Gamma(k+n/\beta)/\Gamma(k)$. Then, the nth moment of X can be determined from (2.8) as

$$\mu'_n = E(X^n) = \mu^n \sum_{m,i=0}^{\infty} p_{m,i}\,\frac{\Gamma\left(3[m+1]/2+i+n/2\right)}{\Gamma\left(3[m+1]/2+i\right)}. \tag{2.18}$$

Some of the most important features and characteristics of a distribution can be studied through moments (e.g., tendency, dispersion, skewness and kurtosis). The central moments and cumulants of X can be determined from the ordinary moments in (2.18) using well-known formulae.

Corollary 2 The nth incomplete of X can be expressed as

$$M_n(s) = \int_0^s x^n\,f(x;\mu,\sigma,\nu)\,dx = \sum_{m,i=0}^{\infty} p_{m,i}\,I_n(s;\mu,k^*,2).$$

Proof Corollary 2:

The nth incomplete moment $I_n(s;\mu,k,2) = \int_0^s x^n\,\pi(x;\mu,k,2)dx$ of $\pi(x;\mu,k,2)$ is easily found by transforming variables $z = (t/\mu)^2$ as

$$I_n(s;\mu,k,2) = \frac{\mu^n}{\Gamma(k)}\,\gamma(n/2+k,(s/\mu)^2).$$

Hence, the nth incomplete of X follows from Equation (2.8) as

$$M_n(s) = \int_0^s x^n\,f(x;\mu,\sigma,\nu)\,dx = \sum_{m,i=0}^{\infty} p_{m,i}\,I_n(s;\mu,k^*,2).$$

The first incomplete moment $M_1(s)$ plays an important role for measuring inequality such as the mean deviations and Lorenz and Bonferroni curves. First, the mean deviations about the mean $\tau_1' = E(X)$ and about the median m of X are determined from $\delta_1 = 2\tau_1' F(\tau_1') - 2M_1(\tau_1')$ and $\delta_2 = \tau_1' - 2M_1(m)$, where $F(\tau_1')$ and $F(m)$ are easily calculated from (2.4).

Another application of $M_1(s)$ refers to the Bonferroni and Lorenz curves of X. These curves are very useful in economics, reliability, demography, insurance and medicine. For a given probability π, the Bonferroni and Lorenz curves are given by $B(\pi) = M_1(p)/(p\tau_1')$ and $L(p) = M_1(p)/\tau_1'$, where $p = Q(\pi) = F^{-1}(\pi)$ can be computed from (2.6).

Corollary 3 The moment generating function (mgf) of X can be reduced to

$$M(s) = \sum_{m,i=0}^{\infty} p_{m,i}\; M_{\mu,k^*,2}(s),$$

where $M_{\mu,k^*,2}(s)$ is the mgf of $\pi(x;\mu,k^*,2)$.

Proof Corollary 3:

The mgf of $\pi(x;\mu,k^*,2)$ follows from Cordeiro *et al.* (2011) as

$$M_{\mu,k^*,2}(s) = \frac{1}{\Gamma(k^*)}\, {}_1\Psi_0 \left[\begin{array}{c} (1,1/2) \\ - \end{array} ; \mu\, s \right], \tag{2.19}$$

where ${}_1\Psi_0$ is the Wright generalized hypergeometric function defined by

$${}_p\Psi_q \left[\begin{array}{c} (\mu_1, A_1), \cdots, (\mu_p, A_p) \\ (\beta_1, B_1), \cdots, (\beta_q, B_q) \end{array} ; x \right] = \sum_{m=0}^{\infty} \frac{\prod_{j=1}^{p} \Gamma(\mu_j + A_j\, m)}{\prod_{j=1}^{q} \Gamma(\beta_j + B_j\, m)} \frac{x^m}{m!}.$$

Hence, the mgf of X can be determined from (2.8) and (2.19) as

$$M(s) = E(\mathrm{e}^{\mathrm{sX}}) = {}_{1\ 0} \left[\begin{array}{c} (1,1/2) \\ - \end{array} ; \mu\, \mathrm{s} \right] = \sum_{\mathrm{m,i}=0}^{\infty} \frac{\mathrm{p_{m,i}}}{\big(3[\mathrm{m}+1]/2 + \mathrm{i}\big)}.$$

2.3 The GOLLMax regression model

Regression analysis involves specifications of the distribution of X given a vector $\mathbf{v} = (v_1, \ldots, v_p)^T$ of explanatory variables. In this section, we adopt systematic components for the three parameters in density (2.5) to allow them varying across the observations (for $i = 1, \ldots, n$) as

$$g_1(\mu_i) = \mathbf{v}_i^T \boldsymbol{\beta}_1, \qquad g_2(\sigma_i) = \mathbf{v}_i^T \boldsymbol{\beta}_2, \qquad g_3(\nu_i) = \mathbf{v}_i^T \boldsymbol{\beta}_3, \tag{2.20}$$

where $g_k : [0,\infty) \to \mathrm{R}$ for $k = 1,2,3$ are known one-to-one link functions continuously twice differentiables, $\mathbf{v}_i^T = (v_{i1}, \ldots, v_{ip})$ is a vector of known explanatory variables for the ith observation, and $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p})^T$, $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2p})^T$ and $\boldsymbol{\beta}_3 = (\beta_{31}, \ldots, \beta_{3p})^T$ are parameter vectors of dimension p. Then, $g_1(\boldsymbol{\mu}) = \mathbf{V}\boldsymbol{\beta}_1$, $g_2(\boldsymbol{\sigma}) = \mathbf{V}\boldsymbol{\beta}_2$, $g_3(\boldsymbol{\nu}) = \mathbf{V}\boldsymbol{\beta}_3$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)^T$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^T$, and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)^T$ is a specified $n \times p$ matrix of full column rank $p < n$. It is assumed that $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are functionally independent. The GOLLMax regression model aims to select the explanatory variables in $\mathbf{V}$ which model μ, σ and ν.

Consider a sample $(x_1, \mathbf{v}_1), \ldots, (x_n, \mathbf{v}_n)$ of n independent observations. Conventional likelihood estimation techniques can be applied here. The total log-likelihood function for the vector of parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T)^T$ from model (2.20) takes the form

$$\begin{aligned} l(\boldsymbol{\psi}) &= n \log\left(\frac{4}{\sqrt{\pi}}\right) + \sum_{i=1}^{n} \log(\nu_i) + \sum_{i=1}^{n} \log(\sigma_i) + \sum_{i=1}^{n} \log\left(\frac{x_i^2}{\mu_i^3}\right) - \sum_{i=1}^{n} \left(\frac{x_i^2}{\mu_i^2}\right) + \\ & (\sigma_i\, \nu_i - 1) \sum_{i=1}^{n} \log\left[\gamma_1(3/2, x_i^2/\mu_i^2)\right] + (\nu_i - 1) \sum_{i=1}^{n} \log\left[1 - \gamma_1(3/2, x_i^2/\mu_i^2)\right] - \\ & 2 \sum_{i=1}^{n} \log\left\{\gamma_1^{\sigma_i\, \nu_i}(3/2, x_i^2/\mu_i^2) + \left[1 - \gamma_1^{\sigma_i}(3/2, x_i^2/\mu_i^2)\right]^{\nu_i}\right\}. \end{aligned} \tag{2.21}$$

The MLE $\widehat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ can be calculated by maximizing the log-likelihood (2.21). The numerical maximization of (2.21) can be done in the `optim` and `gamlss` packages in `R`. Further, we can construct likelihood ratio (LR) statistics for comparing some sub-models with the GOLLMax regression model in the classical way.

2.4 Simulation study for the regression model

We study the performance of the MLEs in the GOLLMax regression model based on $1,000$ replications from the true parameters $\beta_{10} = 2.45$, $\beta_{11} = -0.35$, $\beta_{20} = 0.15$, $\beta_{21} = 0.50$, $\beta_{30} = -0.55$ and $\beta_{31} = 0.20$ for different sample sizes ($n = 100, 350, 850$), using `optim` package in `R` software by BFGS method. We consider three different scenarios for the systematic components:

- scenario 1: $\log(\mu_i) = \beta_{10} + \beta_{11} v_{1i}, \quad \log(\sigma_i) = \beta_{20} + \beta_{21} v_{1i}, \quad \log(\nu_i) = \beta_{30} + \beta_{31} v_{1i}.$
- scenario 2: $\log(\mu_i) = \beta_{10} + \beta_{11} v_{2i}, \quad \log(\sigma_i) = \beta_{20} + \beta_{21} v_{2i}, \quad \log(\nu_i) = \beta_{30} + \beta_{31} v_{2i}.$
- scenario 3: $\log(\mu_i) = \beta_{10} + \beta_{11} v_{1i}, \quad \log(\sigma_i) = \beta_{20} + \beta_{21} v_{2i}, \quad \log(\nu_i) = \beta_{30} + \beta_{31} v_{1i},$

where $v_{1i} \sim$ Binomial$(1, 0.5)$ by considering two groups (0 and 1) and $v_{2i} \sim$ Normal$(0, 0.5)$, for $i = 1, \ldots, n$.

The response variables $x_1, \ldots, x_n$ are generated from the GOLLMax regression model (2.20) as follow:

i. Generate v_{1i} and v_{2i}.

ii. Estimate μ_i, σ_i and ν_i for the fixed scenario.

iii. Generate $u_i \sim$ U(0,1).

iv. Use the steps i., ii. and iii. to calculate the observations x_i's from (2.6).

Tables 2.1, 2.2 and 2.3 gives the average estimates (AEs), biases, mean squared errors (MSEs) of the MLEs, their average lengths (ALs) and we present the empirical coverage probabilities (CPs), say C($\boldsymbol{\psi}$), corresponding to the 95% confidence intervals calculated from the simulations for the parameters $\boldsymbol{\psi}$=$(\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}, \beta_{30}, \beta_{31})^T$. We verify that the AEs tend to be closer to the true parameter values and the MSEs, as well as biases of the sample estimates, decay toward zero when the sample size n increases as expected under first-order asymptotic theory. The figures in Tables 2.1, 2.2 and 2.3 indicate that the CPs are close to the nominal and ALs decrease substantially when n increases, respectively.

Table 2.1. AEs, Biases, MSEs, ALs and CPs of the parameters for the GOLLMax regression model for scenario 1.

n	$\boldsymbol{\psi}$	AEs	Biases	MSEs	ALs	C($\boldsymbol{\psi}$)
	β_{10}	2.437	-0.013	0.174	1.495	0.898
	β_{11}	-0.314	0.036	0.373	2.334	0.981
100	β_{20}	0.353	0.203	0.847	1.355	0.891
	β_{21}	0.484	-0.016	1.855	5.146	0.918
	β_{30}	-0.636	-0.086	0.413	2.354	0.909
	β_{31}	0.245	0.045	0.817	3.417	0.947
	β_{10}	2.479	0.029	0.067	0.883	0.922
	β_{11}	-0.333	0.017	0.153	1.272	0.970
350	β_{20}	0.140	-0.009	0.320	2.049	0.910
	β_{21}	0.472	-0.028	0.790	3.071	0.913
	β_{30}	-0.525	0.025	0.167	1.454	0.911
	β_{31}	0.224	0.024	0.355	2.035	0.928
	β_{10}	2.452	0.002	0.021	0.545	0.937
	β_{11}	-0.344	0.005	0.044	0.740	0.967
850	β_{20}	0.168	0.018	0.126	1.355	0.935
	β_{21}	0.490	-0.009	0.293	1.995	0.949
	β_{30}	-0.557	-0.007	0.062	0.949	0.930
	β_{31}	0.211	0.011	0.124	1.297	0.952

Table 2.2. AEs, Biases, MSEs, ALs and CPs of the parameters for the GOLLMax regression model for scenario 2.

n	$\boldsymbol{\psi}$	AEs	Biases	MSEs	ALs	C($\boldsymbol{\psi}$)
	β_{10}	2.450	0.000	0.111	1.042	0.880
	β_{11}	-0.297	0.053	0.156	1.205	0.784
100	β_{20}	0.300	0.150	0.532	2.387	0.833
	β_{21}	0.410	-0.090	0.952	2.997	0.769
	β_{30}	-0.604	-0.054	0.258	1.686	0.853
	β_{31}	0.301	0.101	0.458	2.084	0.778
	β_{10}	2.460	0.010	0.030	0.601	0.919
	β_{11}	-0.322	0.028	0.058	0.793	0.861
350	β_{20}	0.164	0.014	0.169	1.459	0.913
	β_{21}	0.453	-0.047	0.357	1.991	0.843
	β_{30}	-0.547	0.003	0.082	1.024	0.916
	β_{31}	0.254	0.054	0.176	1.381	0.848
	β_{10}	2.453	0.003	0.011	0.387	0.935
	β_{11}	-0.326	0.024	0.028	0.569	0.888
850	β_{20}	0.157	0.007	0.068	0.971	0.929
	β_{21}	0.451	-0.049	0.184	1.475	0.879
	β_{30}	-0.550	0.000	0.033	0.677	0.931
	β_{31}	0.241	0.041	0.088	1.013	0.876

Due to the difficulty of working analytically with the proposed model, the regularity conditions are verified on the basis of the qq-plots of the sample estimates. The Figures (2.4-2.9), for (n= 850),respectively, are presented to better visualize and understand the behavior of the asymptotic distribution of the MLEs. These plots reveal empirically that the asymptotic distributions of the MLEs tend to the normal distribution (as expected) when the sample size increases. This fact supports that the asymptotic normal distribution provides an adequate approximation to the finite sample distribution of the estimates.

Table 2.3. AEs, Biases, MSEs, ALs and CPs of the parameters for the GOLLMax regression model for scenario 3.

n	$\boldsymbol{\psi}$	AEs	Biases	MSEs	ALs	C($\boldsymbol{\psi}$)
	β_{10}	2.431	-0.019	0.120	1.148	0.881
	β_{11}	-0.337	0.013	0.014	0.430	0.923
100	β_{20}	0.316	0.166	0.614	2.624	0.855
	β_{21}	0.546	0.046	0.102	1.107	0.906
	β_{30}	-0.610	-0.060	0.302	1.878	0.871
	β_{31}	0.215	0.015	0.047	0.757	0.910
	β_{10}	2.436	-0.014	0.027	0.607	0.921
	β_{11}	-0.347	0.003	0.004	0.229	0.947
350	β_{20}	0.213	0.063	0.162	1.493	0.914
	β_{21}	0.509	0.009	0.025	0.571	0.923
	β_{30}	-0.581	-0.031	0.082	1.057	0.915
	β_{31}	0.204	0.004	0.012	0.397	0.937
	β_{10}	2.446	-0.004	0.010	0.387	0.933
	β_{11}	-0.350	0.000	0.001	0.147	0.947
850	β_{20}	0.173	0.023	0.061	0.973	0.936
	β_{21}	0.506	0.006	0.009	0.363	0.924
	β_{30}	-0.559	-0.009	0.032	0.685	0.931
	β_{31}	0.199	-0.001	0.004	0.254	0.951

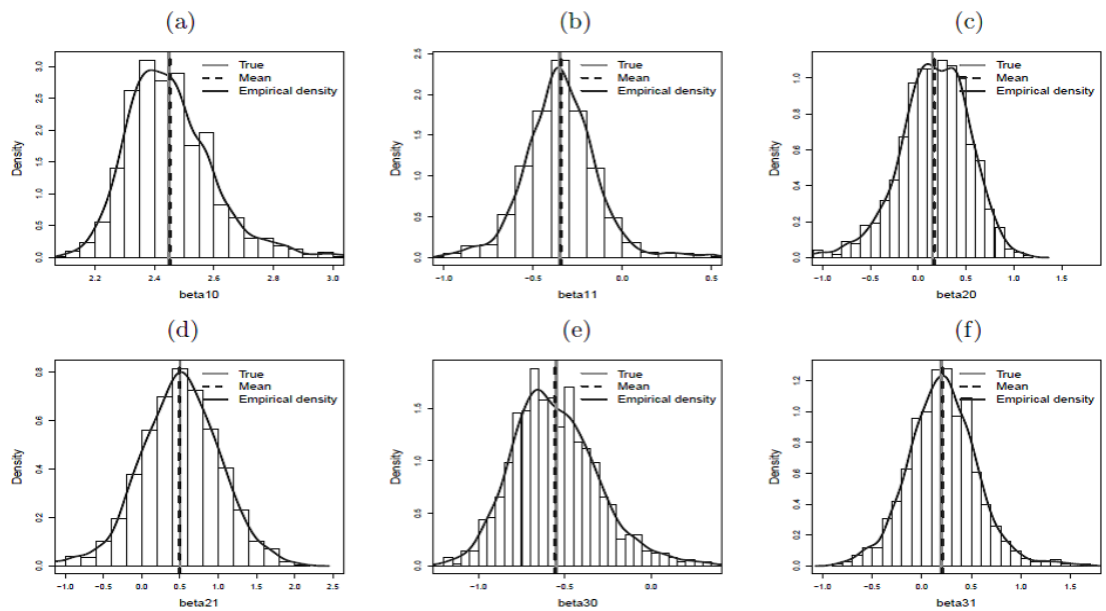

Figure 2.4. Plots of the empirical distributions of the 1,000 parameter estimates for $n = 850$ for scenario 1. (a) For $\hat{\beta}_{10}$. (b) For $\hat{\beta}_{11}$. (c) For $\hat{\beta}_{20}$. (d) For $\hat{\beta}_{21}$. (e) For $\hat{\beta}_{30}$. (f) For $\hat{\beta}_{31}$.

2.5 Diagnostic and residual analysis

We adopt diagnostic measures based on case deletion (global influence) to detect influential observations in the proposed regression model. The case-deletion model with systematic (2.20) is

$$g_1(\mu_l) = \mathbf{v}_l^T \boldsymbol{\beta}_1, \qquad g_2(\sigma_l) = \mathbf{v}_l^T \boldsymbol{\beta}_2, \qquad g_3(\nu_l) = \mathbf{v}_l^T \boldsymbol{\beta}_3 \quad l = 1, \ldots, n, \quad l \neq i. \tag{2.22}$$

In the following, a quantity with subscript "(i)" means the original quantity with the ith observation deleted. For model (2.22), the log-likelihood function for $\boldsymbol{\psi}$ is denoted by $l_{(i)}(\boldsymbol{\psi})$. Let $\widehat{\boldsymbol{\psi}}_{(i)} =$

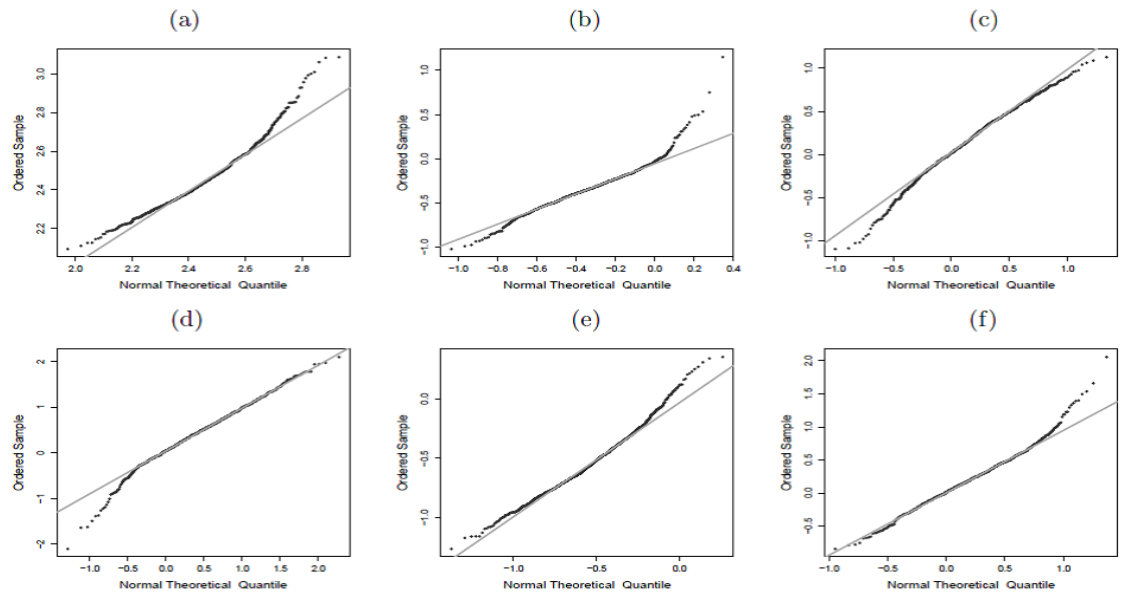

Figure 2.5. Normal probability plots when $n = 850$ for scenario 1. (a) For $\hat{\beta}_{10}$. (b) For $\hat{\beta}_{11}$. (c) For $\hat{\beta}_{20}$. (d) For $\hat{\beta}_{21}$. (e) For $\hat{\beta}_{30}$. (f) For $\hat{\beta}_{31}$.

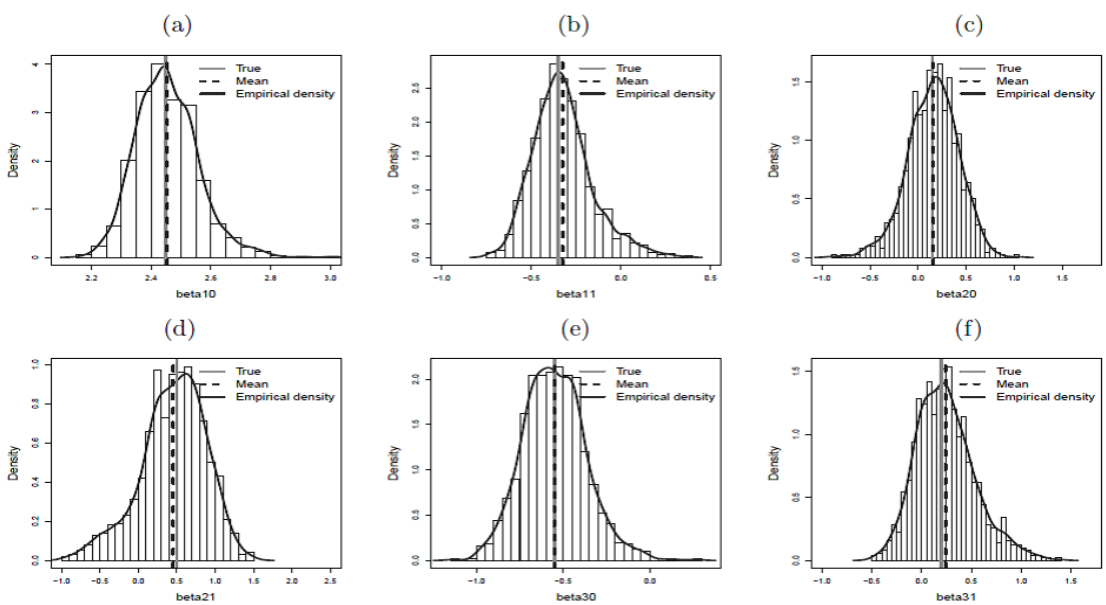

Figure 2.6. Plots of the empirical distributions of the 1,000 parameter estimates for $n = 850$ for scenario 1. (a) For $\hat{\beta}_{10}$. (b) For $\hat{\beta}_{11}$. (c) For $\hat{\beta}_{20}$. (d) For $\hat{\beta}_{21}$. (e) For $\hat{\beta}_{30}$. (f) For $\hat{\beta}_{31}$.

$(\hat{\boldsymbol{\beta}}_{1(i)}^T, \hat{\boldsymbol{\beta}}_{2(i)}^T, \hat{\boldsymbol{\beta}}_{3(i)}^T)^T$ be the MLE of $\boldsymbol{\psi}$ from $l_{(i)}(\boldsymbol{\psi})$. To assess the influence of the ith observation on the MLEs $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T, \hat{\boldsymbol{\beta}}_3^T)^T$, we can compare the difference between $\hat{\boldsymbol{\psi}}_{(i)}$ and $\hat{\boldsymbol{\psi}}$. If deletion of an observation seriously influences the estimates, more attention should be paid to that observation. Hence, if $\hat{\boldsymbol{\psi}}_{(i)}$ is far from $\hat{\boldsymbol{\psi}}$, then the ith observation can be regarded as influential. A first measure of the global influence is defined as the standardized norm of $\hat{\boldsymbol{\psi}}_{(i)} - \hat{\boldsymbol{\psi}}$ (generalized Cook distance), namely

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

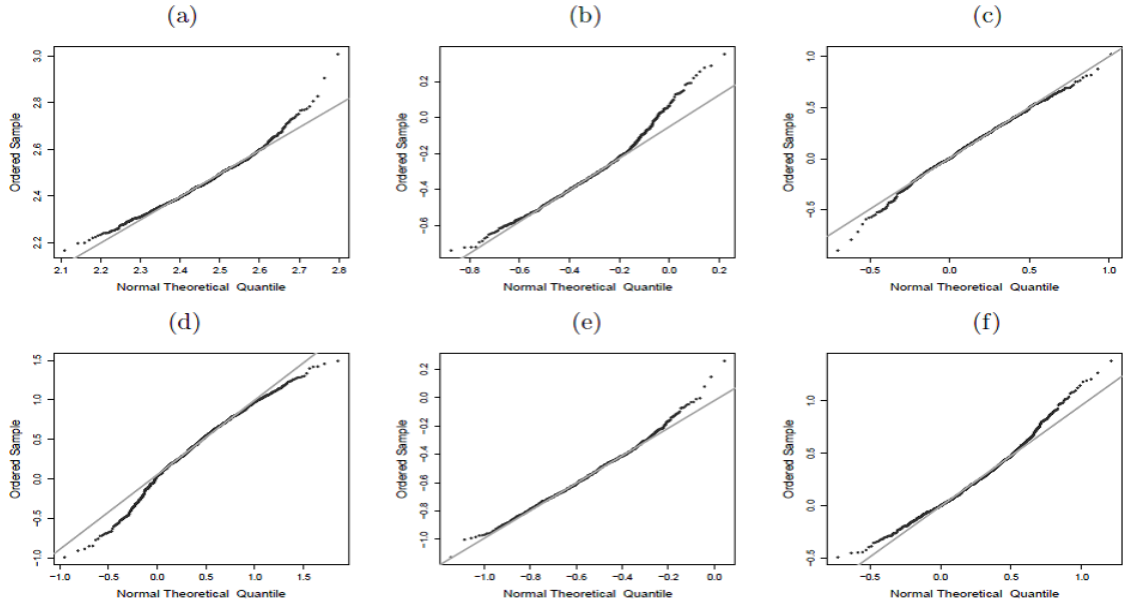

Figure 2.7. Normal probability plots when $n = 850$ for scenario 1. (a) For $\hat{\beta}_{10}$. (b) For $\hat{\beta}_{11}$. (c) For $\hat{\beta}_{20}$. (d) For $\hat{\beta}_{21}$. (e) For $\hat{\beta}_{30}$. (f) For $\hat{\beta}_{31}$.

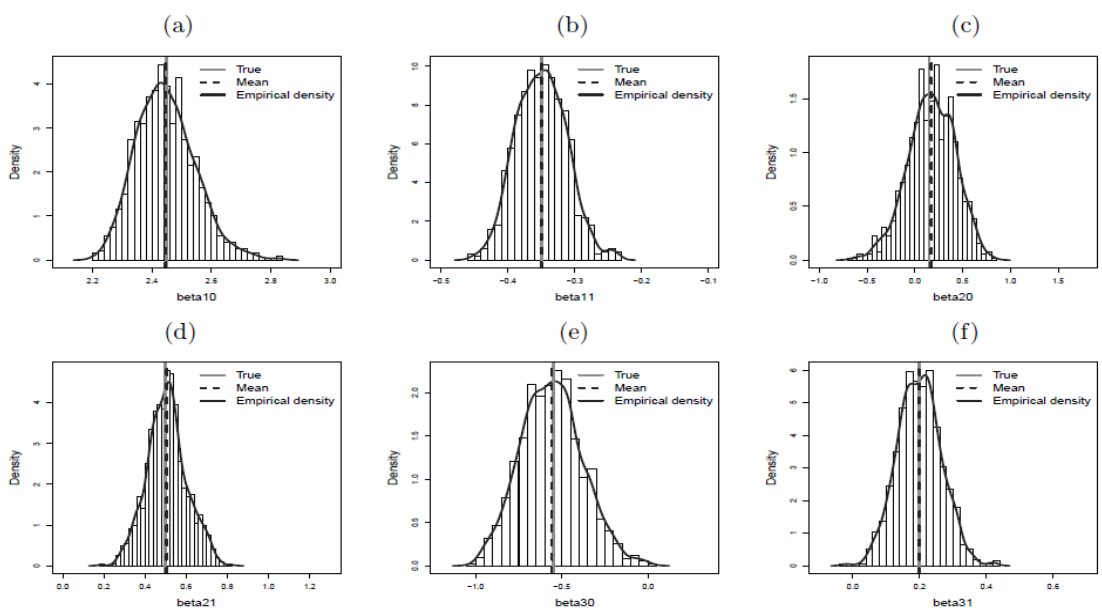

Figure 2.8. Plots of the empirical distributions of the 1,000 parameter estimates for $n = 850$ for scenario 1. (a) For $\hat{\beta}_{10}$. (b) For $\hat{\beta}_{11}$. (c) For $\hat{\beta}_{20}$. (d) For $\hat{\beta}_{21}$. (e) For $\hat{\beta}_{30}$. (f) For $\hat{\beta}_{31}$.

Another popular measure of the difference between $\hat{\boldsymbol{\psi}}_{(i)}$ and $\hat{\boldsymbol{\psi}}$ is the likelihood distance given by

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

We can study departures from the error assumption as well as the presence of outliers for various residuals introduced in the literature but we consider the *quantile residuals* (qr's). For the new regression

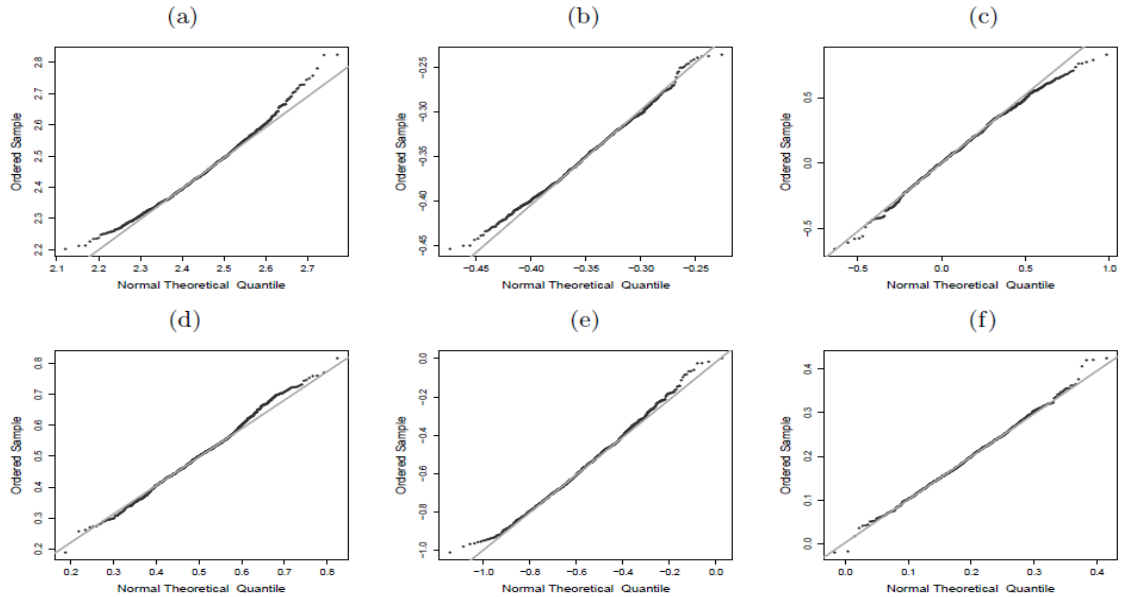

Figure 2.9. Normal probability plots when $n = 850$ for scenario 1. (a) For $\hat{\beta}_{10}$. (b) For $\hat{\beta}_{11}$. (c) For $\hat{\beta}_{20}$. (d) For $\hat{\beta}_{21}$. (e) For $\hat{\beta}_{30}$. (f) For $\hat{\beta}_{31}$.

model, they are defined by

$$\hat{qr}_i = \Phi^{-1}\left\{\frac{\gamma_1^{\hat{\sigma}_i\hat{\nu}_i}(3/2, x_i^2/\hat{\mu}_i^2)}{\gamma_1^{\hat{\sigma}_i\hat{\nu}_i}(3/2, x_i^2/\hat{\mu}_i^2) + \left[1 - \gamma_1^{\hat{\sigma}_i}(3/2, x_i^2/\hat{\mu}_i^2)\right]^{\hat{\nu}_i}}\right\}, \tag{2.23}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative standard normal distribution.

We built envelopes to enable better interpretation of the probability normal plot of the residuals. These envelopes are simulated confidence bands described by Atkinson (1985) that contain the residuals, such that if the model is well-fitted, the majority of points will be randomly distributed within these bands.

2.6 Applications

In this section, we perform three applications in the engineering area and analyze three real data sets. In the first and second applications, we fit the Maxwell, EMax, OLLGMax and GOLLMax (nested models). In addition, we analyze the data sets by using non-nested alternative models. We consider the Weibull distribution with scale parameter $\mu > 0$ and shape parameter $\delta > 0$.

The cdf of the gamma-Maxwell (GMax) distribution (Iriarte *et al.*, 2016) is given by

$$F(x) = \frac{\gamma\left(\delta, -\log\left(1 - \frac{2}{\pi}\gamma\left(\frac{3}{2}, \mu x^2\right)\right)\right)}{\Gamma(\delta)}, \quad x > 0,$$

where $\mu > 0$ is a scale parameter and $\delta > 0$ is a shape parameter.

The cdf of the three-parameter transmuted exponentiated Maxwell (TEMax) distribution (Vanegas *et al.*, 2017) is given by

$$F(x) = (1 - \lambda)\, G(x)^{\delta} - \lambda\, G(x)^{2\,\delta}, \quad x > 0,$$

where $\mu > 0$ is a scale parameter, $\delta > 0$ is a shape parameter, $|\lambda| \leq 1$ is a parameter that makes the asymmetry more flexible and $G(x) = G(x;\mu)$ is the Maxwell cdf given by (2.1).

In the third application, we illustrate the flexibility of the GOLLMax regression model.

In the applications, we determine the MLEs and their estimated standard errors (SEs) (given in parentheses) of the model parameters and the values of the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Cramér-von Mises (W^*), Anderson Darling (A^*) and Kolmogorov-Smirnov (KS) statistics for the fitted models. The lower the values of these measures, the better the fit. For all distributions, the parameters are estimated by maximum likelihood. We adopt the `AdequacyModel` script BFGS and CG algorithms for the first two applications, whereas the `gamlss` function by RS method described by (Rigby and Stasinopoulos, 2005) in the `R` software is used for the third application.

2.6.1 Application 1: Strength data

The data set with 49 observations presented by Chen (2006) was obtained from a process of manufacturing a plastic laminate whose resistance must exceed a few pounds per square inch (psi). We calculate the MLEs of the model parameters and the above statistics for each fitted model to these data. The results are presented in Table 2.4. The five statistical measures are favorable to the GOLLMax distribution, which can be chosen as the best distribution to explain the current data.

Table 2.4. MLEs of the model parameters, the their SEs (given in parentheses) and the statistics: AIC, BIC, W^*, A^* and KS (p-value associated in parentheses) for the strength data.

Model	μ	σ	ν	AIC	BIC	W^*	A^*	KS
GOLLMax	21.4169	18.8193	0.3115	411.9	417.6	0.0382	0.2556	0.0854
	(0.0002)	(0.0017)	(0.0375)					(0.8364)
OLLMax	40.6034	1	1.2166	417.5	421.3	0.1202	0.7397	0.0949
	(2.0973)	(-)	(0.1534)					(0.7327)
EMax	36.3324	1.5610	1	416.2	420.0	0.1140	0.7000	0.1370
	(2.7689)	(0.3508)	(-)					(0.2892)
Maxwell	40.9888	1	1	418.0	419.9	0.1214	0.7448	0.1335
	(2.3905)	(-)	(-)					(0.3179)
Model	μ	δ	λ	AIC	BIC	W^*	A^*	KS
GMax	34.5788	1.4853		416.6	420.4	0.1195	0.7331	0.1339
	(3.5018)	(0.3053)						(0.3147)
TEMax	38.6766	1.6259	0.4090	417.3	422.9	0.1005	0.6181	0.1315
	(0.0006)	(0.2806)	(0.2932)					(0.3347)
Weibull	52.9449	2.8904		420.1	423.9	0.1587	0.9680	0.1145
	(2.7770)	(0.3039)						(0.5045)

Formal tests to verify the inclusion or not of the additional parameters σ and ν in the proposed distribution can be done based on the LR statistics as listed in Table 2.5. We reject the null hypotheses in the three tests in favor of the GOLLMax distribution. The rejection is significant and provides clear evidence of the flexibility of the shape parameters σ and ν when modeling real data with bimodal characteristics. More information is provided by a visual comparison of the data histogram and fitted density functions. The plots of the fitted GOLLMax, TEMax and Weibull densities are displayed in Figure 2.10a. The estimated GOLLMax density provides the closest fit to the histogram of the data.

In order to assess if the model is appropriate, the plots of the fitted GOLLMax, TEMax and Weibull cumulative distributions and the empirical cdf are displayed in Figure 2.10b. They also indicate that the wider distribution provides a good fit to these data.

Table 2.5. LR tests for strength data.

Models	Hypotheses	Statistic w	p-value
GOLLMax vs OLLMax	$H_0 : \sigma = 1$ vs $H_1 : H_0$ is false	7.5	0.0059
GOLLMax vs EMax	$H_0 : \nu = 1$ vs $H_1 : H_0$ is false	6.3	0.0120
GOLLMax vs Maxwell	$H_0 : \sigma = \nu = 1$ vs $H_1 : H_0$ is false	10.0	0.0064

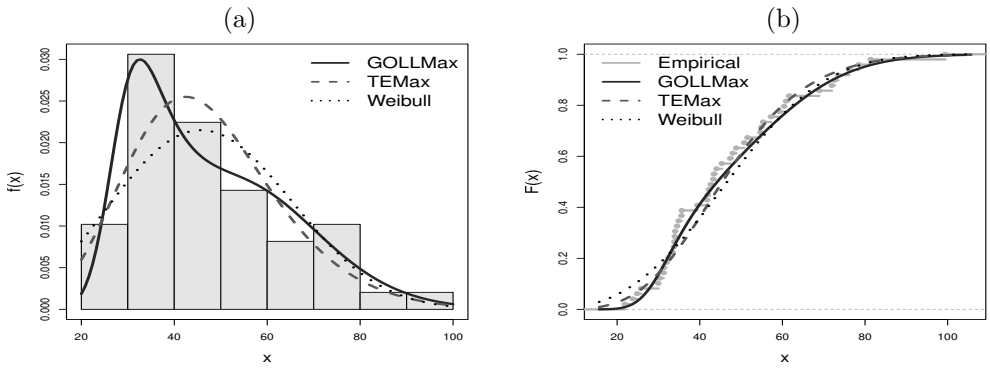

Figure 2.10. (a) Estimated densities of the GOLLMax, TEMax and Weibull models for strength data. (b) Estimated cumulative functions of the GOLLMax, TEMax and Weibull models for strength data.

2.6.2 Application 2: Image data

The data set was extracted from an image of Foulum (Denmark) obtained by the EMISAR sensor (Lee and Pottier (2009)) jointly built by the Electro Magnetics Institute (EMI), the Technical University of Denmark (TUD), and its Danish Centre for Remote Sensing (DCRS), operated at C- and L-bands (though not simultaneously) with quad-polarizations. The data are obtained at http://earth.eo.esa.int /polsarpro/datasets.html by means of the PolSARpro software and, for each geographic position, each one of its element consists in norm squared of a complex number, which represents the information of the polarization channel resulting of a pulse both transmitted and recorded in horizontal direction. A scenario of this data set is presented in Alizadeh *et al.* (2017).

We calculate the MLEs of the model parameters and the above statistics for the fitted models to these data. The results are reported in Table 2.6. The five statistics agree with the suitability of the proposed model. In fact, the lowest values of them indicate that the GOLLMax distribution could be chosen as the best model to these data.

Formal tests to verify the inclusion or not of the additional parameters σ and ν in the proposed distribution can be performed based on the LR statistics given in Table 2.7. We reject the null hypotheses in the three tests in favor of the GOLLMax distribution. The rejection is significant and provides clear evidence of the flexibility of the shape parameters σ and ν for modeling real data with bimodal characteristics. More information is provided by a visual comparison of the data histogram and adjusted densities. Figure 2.11a displays the plots of the estimated GOLLMax, OLLMax and Weibull densities. The estimated GOLLMax density provides the closest fit to the histogram of these data.

In order to assess if the model is appropriate, the plots of the estimated cdfs of the GOLLMax, OLLMax and Weibull distributions and the empirical cdf are displayed in Figure 2.11b. They also support that the GOLLMax distribution provides a good fit to these data.

Table 2.6. MLEs of the model parameters, their SEs (given in parentheses) and the AIC, BIC, W^*, A^* and KS (p-value associated in parentheses) statistics for the image data.

Model	μ	σ	ν	AIC	BIC	W^*	A^*	KS
GOLLMax	0.0658	4.6876	0.1843	-243.0	-235.1	0.0642	0.4371	0.0579
	(0.0133)	(2.2051)	(0.0755)					(0.8871)
OLLMax	0.1108	1	0.5178	-233.4	-228.2	0.2089	1.1761	0.1162
	(0.0067)	(-)	(0.0483)					(0.1307)
EMax	0.1701	0.4222	1	-226.9	-221.7	0.2871	1.6149	0.1284
	(0.0122)	(0.0505)	(-)					(0.0714)
Maxwell	0.1290	1	1	-165.4	-162.8	0.2970	1.6711	0.3074
	(0.0052)	(-)	(-)					(<0.0001)
Model	μ	δ	λ	AIC	BIC	W^*	A^*	KS
GMax	0.1879	0.4229		-226.4	-221.2	0.2933	1.6501	0.1292
	(0.0150)	(0.0511)						(0.0685)
TEMax	0.1748	0.4428	0.1585	-225.2	-217.3	0.2851	1.5987	0.1232
	(0.0170)	(0.0613)	(0.3047)					(0.0930)
Weibull	0.1456	1.5051		-226.3	-221.1	0.2945	1.6488	0.1188
	(0.0101)	(0.1198)						(0.1156)

Table 2.7. LR tests for the image data.

Models	Hypotheses	Statistic w	p-value
GOLLMax vs OLLMax	$H_0 : \sigma = 1$ vs $H_1 : H_0$ is false	11.6	0.0006
GOLLMax vs EMax	$H_0 : \nu = 1$ vs $H_1 : H_0$ is false	18.2	0.00001
GOLLMax vs Maxwell	$H_0 : \sigma = \nu = 1$ vs $H_1 : H_0$ is false	81.6	<0.0001

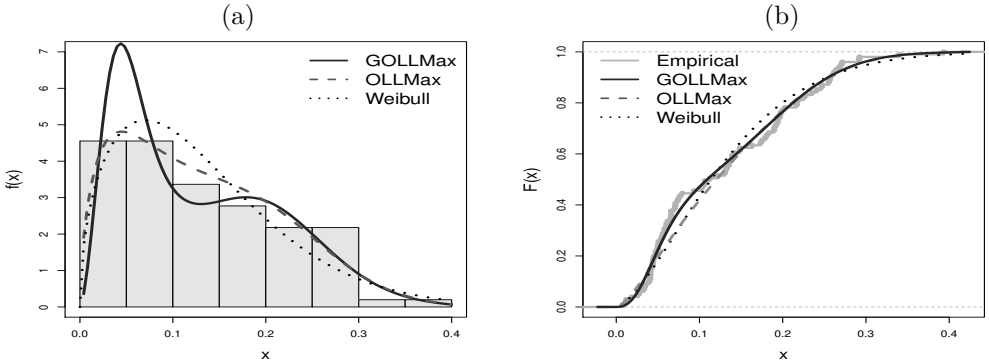

Figure 2.11. (a) Estimated densities of the GOLLMax, OLLMax and Weibull models to the image data. (b) Estimated cumulative functions of the GOLLMax, OLLMax and Weibull models to the image data.

2.6.3 Application 3: Brittle materials

Basu *et al.* (2009) presented a detailed analysis of the data on resistance measurements (in MPa) for different components of ceramic materials and a glass material. In this way, the data set is divided into four subsets with measurements of materials with different chemical compositions, such as, ZrO_2, $ZrO_2 - TiB_2$, Si_3N_4 and glass (with unknown composition). The interest is to verify the strength properties of solid materials, extremely brittle as glass and the most resistant as the materials to be made of ZrO_2 and Si_3N_4.

The Weibull distribution is an alternative in applications involving resistance studies of brittle materials, for example, in Xu et al. (2001). Basu *et al.* (2009) proposed the Weibull distribution

for an alternative modeling and compared it with other two-parameter distributions. However, in the application, the analyzes are performed considering the compositions separately. Here, we propose a joint analysis by considering a regression model with three systematic components given by (2.20).

The explanatory variables are:

- x_i -observed value of the strength of the material (until rupture or crack occurs);
- v_i - chemical compounds of materials with four levels (ZrO_2, $ZrO_2 - TiB_2$, Si_3N_4 and Glass) is defined by dummy variables: ZrO_2 ($v_{i1} = 0$, $v_{i2} = 0$ and $v_{i3} = 0$), $ZrO_2 - TiB_2$ ($v_{i1} = 1$, $v_{i2} = 0$ and $v_{i3} = 0$), Si_3N_4 ($v_{i1} = 0$, $v_{i2} = 1$ and $v_{i3} = 0$) and Glass Si_3N_4 ($v_{i1} = 0$, $v_{i2} = 0$ and $v_{i3} = 1$).

First, we perform an exploratory analysis for these data. We can verify by means of Figure 2.12 that the different groups of chemical compounds have bimodality and asymmetry. This behavior indicates that a more flexible model, for example, the GOLLMax distribution can be more adequate than the most popular gamma, exponential, log-normal and Weibull distributions.

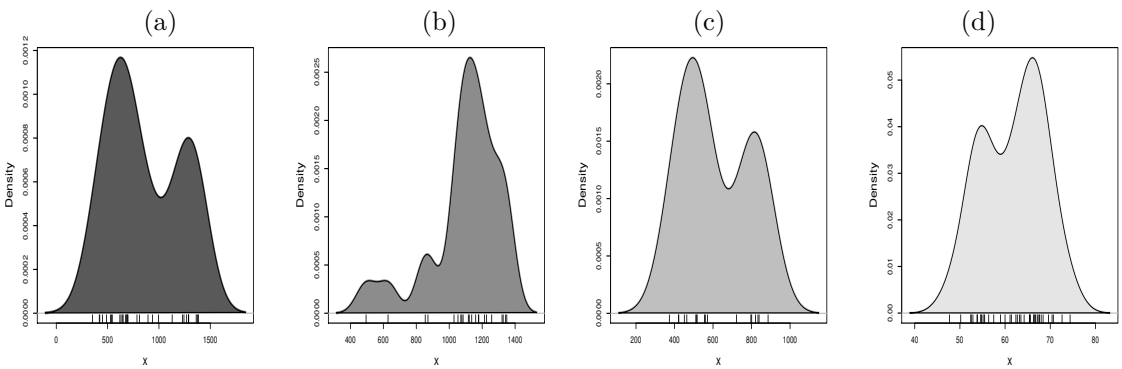

Figure 2.12. Plots for brittle materials data. (a) ZrO2. (b) ZrO2-TiB2. (c) Si3N4. (d) Glass.

Considering only the response variable x_i, we verify the suitability of the proposed model and compare it with the Weibull distribution. Table 2.8 gives the MLEs and their SEs (in parentheses) and the AIC, BIC, W^*, A^* and KS statistics.

Table 2.8. MLEs of the model parameters for brittle materials data, their SEs (given in parentheses) and the AIC, BIC, W^*, A^* and KS statistic.

Model	μ	σ	ν	AIC	BIC	W^*	A^*	KS
GOLLMax	230.0736	3.6791	0.0992	1623.6	1631.8	0.5419	4.0776	0.1521
	(0.0297)	(0.0178)	(0.0075)					(0.0107)
	μ	δ		AIC	BIC	W^*	A^*	KS
Weibull	591.0127	1.0321		1669.5	1674.9	1.0800	7.2024	0.2428
	(56.5923)	(0.0822)						(<0.0001)

Figure 2.13a displays the estimated GOLLMax and Weibull densities and the histogram to verify which model is more appropriate. As an alternative to check the quality fit, Figure 2.13b displays the hrfs of the GOLLMax and Weibull models and the empirical hazard function. We conclude that the GOLLMax distribution provides a better fit to these data.

We consider the following systematic structures:

$$\mu_i = \exp(\beta_{10} + \beta_{11} v_{i1} + \beta_{12} v_{i2} + \beta_{13} v_{i3}), \quad \sigma_i = \exp(\beta_{20} + \beta_{21} v_{i1} + \beta_{22} v_{i2} + \beta_{23} v_{i3}) \quad \text{and}$$

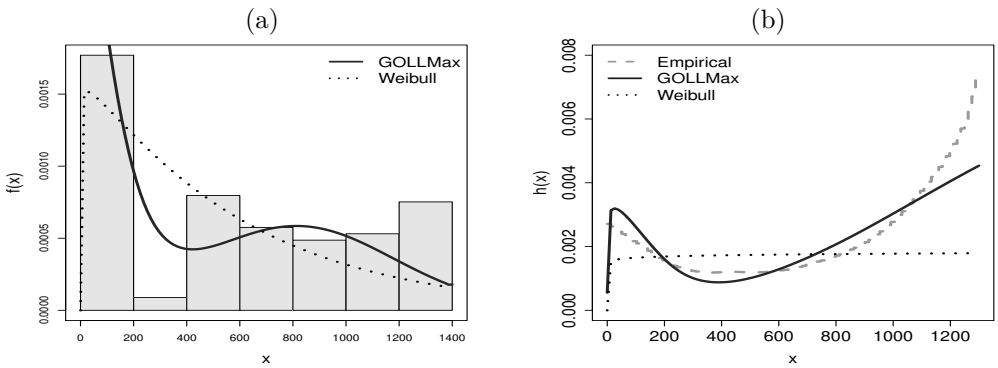

Figure 2.13. (a) Estimated densities of the GOLLMax and Weibull models for brittle materials data. (b) Estimated cumulative functions of the GOLLMax and Weibull models for brittle materials data.

Table 2.9. MLEs, SEs and p-values for the fitted GOLLMax regression model to the brittle materials data.

Parameters	Estimates	SEs	p-values
β_{10}	5.9882	0.0673	< 0.0001
β_{11}	1.0437	0.0816	< 0.0001
β_{12}	-0.4956	0.0857	< 0.0001
β_{13}	-2.1062	0.0695	< 0.0001
β_{20}	2.6242	0.2220	< 0.0001
β_{21}	-2.8636	0.2362	< 0.0001
β_{22}	1.4901	0.3797	0.0001
β_{23}	-2.1661	0.2286	< 0.0001
β_{30}	-1.2495	0.1784	< 0.0001
β_{31}	2.3455	0.2241	< 0.0001
β_{32}	0.0570	0.3064	0.8530
β_{33}	2.4431	0.2597	< 0.0001

$$\nu_i = \exp(\beta_{30} + \beta_{31} v_{i1} + \beta_{32} v_{i2} + \beta_{33} v_{i3}), \quad i = 1, \ldots, 113.$$

In addition, we present the estimates of the parameters, SEs and the associated p-values of the MLEs in Table 2.9. The figures in this table give evidence that the presence of the covariate v_{i2} (Si_3N_4) is not significant at a significance level of 5% in the regression structure for the parameter ν_i in relation to the ZrO_2 level. This fact confirms the exploratory analysis shown in Figure 2.12 in which the groups ZrO_2 in Figure 2.12a and Si_3N_4 in Figure 2.12c present similar bimodal forms thus contributing to the non-significance.

Table 2.9 suggests that the materials $ZrO_2 - TiB_2$, Si_3N_4 and glass are statistically different from the ZrO_2 material in all structures, except as mentioned above. This fact reveals the modeling ability of the proposed structure to model the scale of the data using the parameter μ_i and asymmetry and bimodality through the parameters σ_i and ν_i, respectively.

Model checking

The next step is to detect possible influential points in the GOLLMax regression model. The generalized Cook distance distance and likelihood distance are displayed in Figure 2.14. These plots reveal that the cases 29 and 56 are possible influential observations.

Further, we verify the quality of the adjustment of the GOLLMax regression by constructing the normal probability plot for the qr's for the waste diversion with simulated envelope. There is evidence of a good fit of the GOLLMax regression model as illustrated in Figures 2.15a and 2.15b.

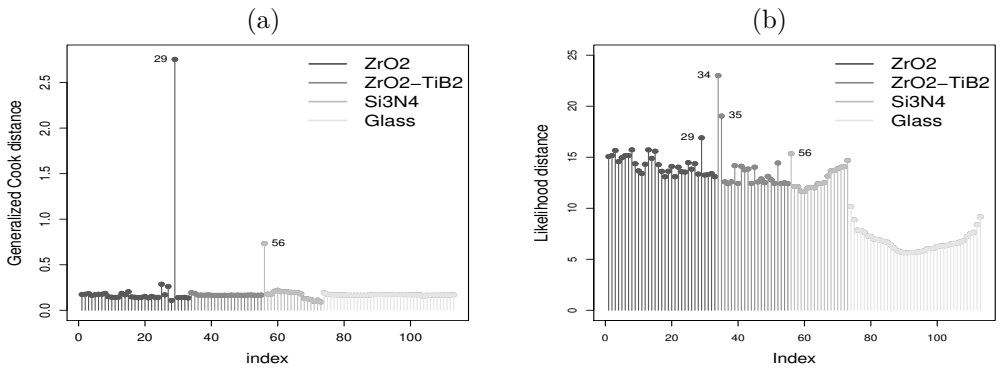

Figure 2.14. (a) Generalized Cook distance for the GOLLMax regression model to the brittle materials data. (b) Likelihood distance for the GOLLMax regression model to the brittle materials data.

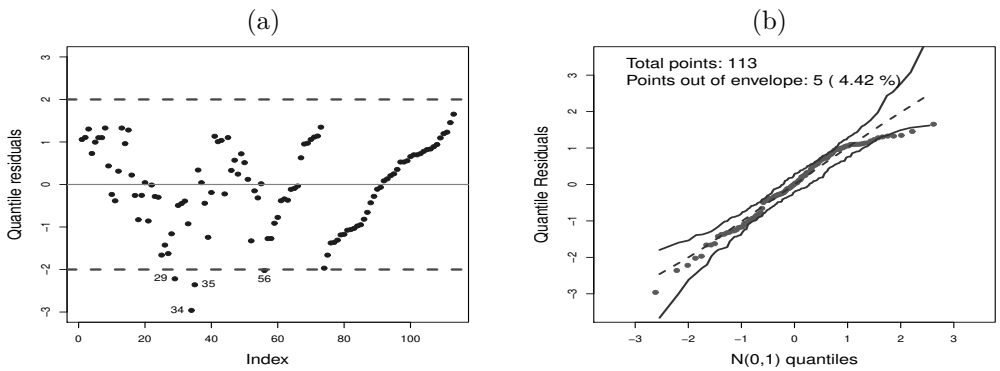

Figure 2.15. (a) Index plot of the qr's from the fitted GOLLMax regression model to brittle materials data. (b) Normal probability plot for the qr's with envelopes.

In this analysis, we make an analogy with the methodology of survival analysis. We consider that the event of interest is the breaking or rupture of the material, that is, applied a force of resistance in MPa how much the material supports until breakage or rupture occurs.

The survival function corresponding to (2.5) is

$$S(x) = 1 - \frac{\gamma_1^{\sigma\nu}(3/2, x^2/\mu^2)}{\gamma_1^{\sigma\nu}(3/2, x^2/\mu^2) + [1 - \gamma_1^{\sigma}(3/2, x^2/\mu^2)]^{\nu}}. \tag{2.24}$$

In Figure 2.16a, we display the Kaplan-Meier empirical curves and estimated survival functions determined from (2.24). Note that glass is the weakest material with least resistance. For a strength of 800 MPa the probabilities that in the materials ZrO_2, $ZrO_2 - TiB_2$ and Si_3N_4 do not occur the event of interest are 0.4912, 0.9446 and 0.1633, respectively. In Figure 2.16b, we give the estimated hrf for each material and verify the presence of different forms for this function. Based on this figure, we note that both models show satisfactory fits. However, the GOLLMax regression model presents a better fit to the current data.

2.7 Concluding Remarks

We propose a three-parameter model called the *generalized odd log-logistic Maxwell* (GOLLMax) distribution, which includes as special cases the odd log-logistic Maxwell (OLLMax), exponentiated

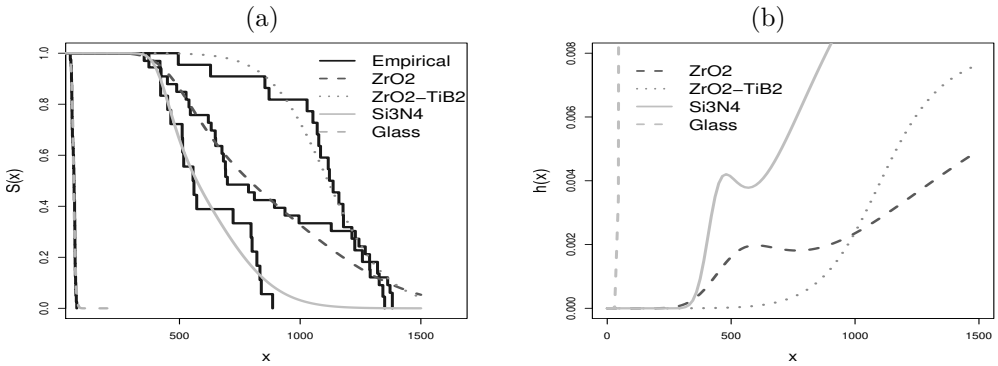

Figure 2.16. The GOLLMax regression model. (a) Estimated survival functions and the empirical survival for the brittle materials data. (b) Estimated hrf for the brittle materials data.

Maxwell (EMax) and Maxwell distributions. We provide some of its mathematical properties. We define a GOLLMax regression model with three systematic components based on the new distribution. The proposed model serves as an important extension to several existing regression models and could be a valuable addition to the literature. The maximum likelihood method is described for estimating the model parameters. Some simulations are performed for different parameter settings and sample sizes to evaluate the precision of the maximum likelihood estimates. Diagnostic analysis is presented to assess global influences. We also discuss the sensitivity of the estimates from the fitted model via quantile residuals. The utility of the introduced models is discussed by means of three real data sets.

References

Alizadeh, M., Cordeiro, G.M., Nascimento, A.D., Lima, M.D.C.S. and Ortega, E.M.M. (2017). Odd-Burr generalized family of distributions with some applications. *Journal of Statistical Computation and Simulation*, **87**, 367-389.

Atkinson, A.C. (1985). *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis*. Clarendon Press, Oxford.

Basu, B., Tiwari, D., Kundu, D. and Prasad, R. (2009). Is Weibull distribution the most appropriate statistical strength distribution for brittle materials?. *Ceramics International*, **35**, 237-246.

Brilliantov, N.V. and Poschel, T. (2000). *In Granular Gases*. Springer, Berlin.

Chen, C. (2006). Tests of fit for the three-parameter log-normal distribution. *Computational statistics and data analysis*, **50**, 1418-1440.

Cordeiro, G.M., Alizadeh, M., Ozel, G., Hosseini, B., Ortega, E.M.M. and Altun, E. (2017). The generalized odd log-logistic family of distributions: properties, regression models and applications. *Journal of Statistical Computation and Simulation*, **87**, 908-932.

Dey, S., Dey, T., Ali, S. and Mulekar, M.S. (2016). Two-parameter Maxwell distribution: Properties and different methods of estimation. *Journal of Statistical Theory and Practice*, **10**, 291-310.

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

Gradshteyn, I.S. and Ryzhik, I.M. (2000). *Table of Integrals, Series and Products*. Academic Press, San Diego.

Iriarte, Y.A., Astorga, J.M., Bolfarine, H. and Gómez, H.W. (2017). Gamma-Maxwell distribution. *Communications in Statistics-Theory and Methods*, **46**, 4264-4274.

Kazmi, S., Aslam, M. and Ali, S. (2012). On the Bayesian estimation for two component mixture of Maxwell distribution, assuming type I censored data. *SOURCE International Journal of Applied Science and Technology*, **2**, 197-218.

Krishna, H. and Malik, M. (2012). Reliability estimation in Maxwell distribution with progressively type-II censored data. *Journal of Statistical Computation and Simulation*, **82**, 623-641.

Lee, J.S. and Pottier, E. (2009). *Polarimetric radar imaging: from basics to applications*. CRC press.

Prigogine, I., Xhrouet, E. (1949). On the perturbation of Maxwell distribution function by chemical reactions in gases. *Physica*, **15**, 913-932.

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

Stacy, E. W. (1962). A generalization of the gamma distribution. *Annals of Mathematical Statistics*, **33**, 1187-1192.

Tomer, S.K. and Panwar, M. S. (2015). Estimation procedures for Maxwell distribution under type-I progressive hybrid censoring scheme. *Journal of Statistical Computation and Simulation*, **85**, 339-356.

Venegas, O., Iriarte, Y.A., Astorga, J.M., Borger, A., Bolfarine, H. and Gómez, H.W. (2017). A New Generalization of the Maxwell Distribution. *Applied Mathematics and Information Sciences*, **11**, 867-876.

Xu, Y., Cheng, L., Zhang, L., Yan, D. and You, C. (2001). Optimization of sample number for Weibull function of brittle materials strength. *Ceramics International*, **27**, 239-241.

Weisberg, S. (2005). *Applied linear regression*. John Wiley and Sons, New York.

3 THE GENERALIZED ODD LOG-LOGISTIC MAXWELL SEMIPARAMETRIC REGRESSION MODEL FOR CENSORED AND UNCENSORED DATA

Abstract: We propose a semiparametric regression model considering the cubic spline for nonlinear effects for censored and uncensored data. Parameter estimates of the generalized odd log-logistic Maxwell model are obtained by penalized maximum likelihood considering nonlinear effects. Some global-influence measurements and quantile residuals are also investigated. Several Monte Carlo simulations are performed for inference purposes under different nonlinear shape configurations, effective degree of freedom and sample sizes. The proposals are illustrated by two applications to real data set.

Keywords: Censored data; Maxwell distribution; Penalized likelihood; Residuals; Semiparametric regression.

3.1 Introduction

Regression models are very useful tools in the statistical analysis of data and can be applied in different areas of knowledge, such as biology, engineering, agriculture and health, among others, is one of the most used techniques to model and study the relationship of one or more explanatory (or covariate) variables and one dependent (or response) variable. The simplest form of these relationships is the linear one, and over many years the normal linear models have been the most used in an attempt to explain such relationships.

However, in many studies the assumption of linearity between the response variable and the explanatory variables. It is not always appropriate. Like, assume normal distribution for data modeling that in many situations present asymmetrical behavior. For example, the relation between body mass index and patients age such measurements obtained preoperatively of the liver transplantation. The *body mass index* (BMI) is an internationally-standard measure to assess whether a person has an ideal weight. It is adopted as a metric by the World Health Organization (WHO) to measure obesity. Overweight and obesity, as indicated by the BMI, are risk factors for diseases like arterial hypertension, coronary artery disease, metabolic syndrome, respiratory diseases, digestive tract diseases, psychiatric disturbances, cancer, osteoarthritis and diabetes mellitus, besides other pathologies considered to be serious public health problems. We consider the data from the Kelly, et. al, (2012).

In this data set we can see in Figure 3.1a the relationship between the response variable ($y =$ BMI) and ($x_1 =$ age) does not have a linear effect, being a behavior closer to a point cloud. We can also verify that the response variable has a certain degree of asymmetry Figure 3.1b.

A second data set refers to the monoclonal gammopathy of undetermined significance (MGUS) data set included in the library `suvival` in `R` software. In Figure 3.2a and b, y denotes the time in days from diagnosis to last follow-up and as covariates, ($x_1 =$ age in years at the detection of MGUS), ($x_2 =$ size of the monoclonal protein spike at diagnosis) and ($x_3 =$ sex, a factor with level Male and reference Female). It can be seen from plots for Figure 3.2 that the problems mentioned in the first data set also apply to these data. In addition, in this data we have the presence of censored data. Thus, the behavior of the empirical risk function is presented in Figure 3.2c.

We propose the generalized odd log-logistic Maxwell semiparametric regression model to solve the problems above. In addition, to solve the issue of nonlinear behavior and in order to obtain a more flexible model for the data we use cubic splines in this work. Thus, the inclusion of cubic splines in

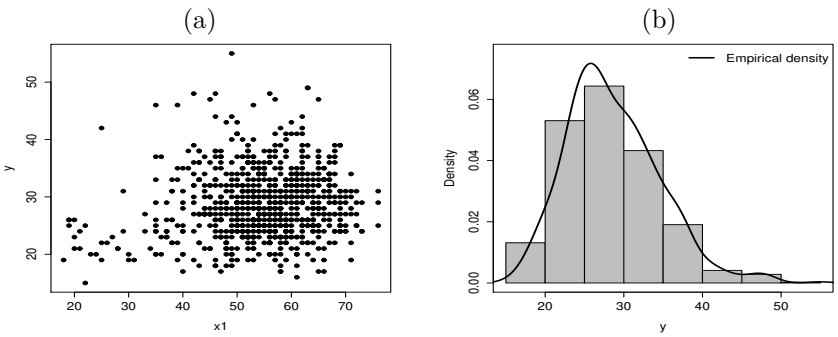

Figure 3.1. (a) Nonlinear effect of BMI versus age. (b) BMI empirical density.

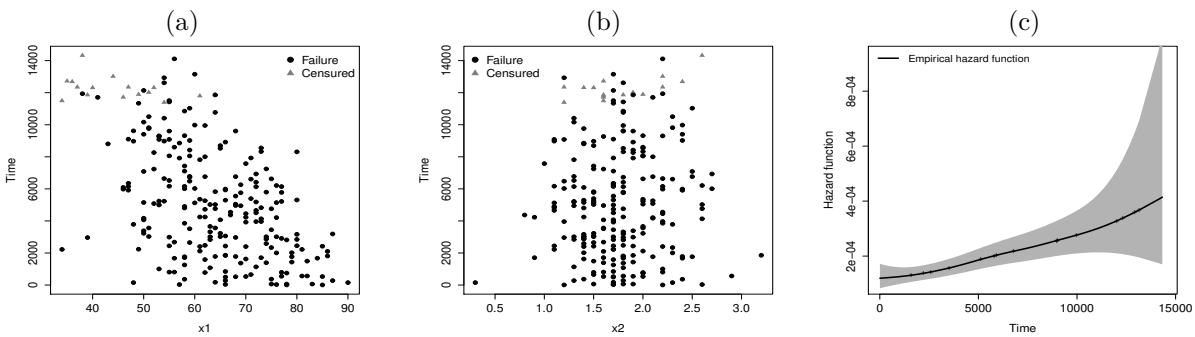

Figure 3.2. (a) Nonlinear effect of time versus age in years. (b) Nonlinear effect of time versus size of monoclonal protein spike at diagnosis. (c) Empirical hazard function.

the model requires, in addition to descriptive and exploratory analyzes, diagnostic analyzes to assess the suitability of the model.

The Monte Carlo simulation studies are conducted to evaluate the performance of this model by means of bias, variance and mean squared error (MSE). For different sample sizes, effective degrees of freedom and effects of nonlinear shapes, simulation studies are performed considering censored and uncensored values. We perform global influence and develop residual analysis based on the quantile residuals.

The chapter is organized as follows. In Section 3.2, we define the *generalized odd log-logistic Maxwell* (GOLLMax) distribution. In Section 3.3, we propose the GOLLMax semiparametric regression model using cubic spline to estimate the nonlinear effects of the covariates. We provide some simulation results in Section 3.4 to verify the asymptotic behavior of the MLEs and the nonlinear effects when the sample size increases. In Section 3.5, we illustrate the flexibility of the proposed regression model by means of two applications. Finally, we offer some conclusions in Section 3.6.

3.2 The GOLLMax distribution

A random variable Y has the GOLLMax distribution if its cumulative distribution function (cdf) and probability density function (pdf) are (for $y > 0$)

$$F(y;\mu,\sigma,\nu) = \frac{\gamma_1^{\nu\,\sigma}(3/2, y^2/\mu^2)}{\gamma_1^{\nu\sigma}(3/2, y^2/\mu^2) + [1-\gamma_1^{\sigma}(3/2, y^2/\mu^2)]^{\nu}} \tag{3.1}$$

and

$$f(y;\mu,\sigma,\nu)=\frac{4\,\nu\,\sigma\,y^2}{\sqrt{\pi}\,\mu^3}\,\exp\left(-\frac{y^2}{\mu^2}\right)\frac{\gamma_1^{\nu\sigma-1}(3/2,y^2/\mu^2)\left[1-\gamma_1^{\sigma}(3/2,y^2/\mu^2)\right]^{\nu-1}}{\left\{\gamma_1^{\nu\sigma}(3/2,y^2/\mu^2)+\left[1-\gamma_1^{\sigma}(3/2,y^2/\mu^2)\right]^{\nu}\right\}^2}, \tag{3.2}$$

respectively, where $\nu>0$ and $\sigma>0$ are two extra shape parameters and $\mu>0$ is scale parameter. The $\gamma_1(p,u)=\gamma(p,u)/\Gamma(p)$ is the incomplete gamma function ratio, $\gamma(p,u)=\int_0^u w^{p-1}\mathrm{e}^{-\mathrm{w}}\mathrm{dw}$ is the incomplete gamma function and $\Gamma(\cdot)$ is the gamma function. Henceforth, if Y is a random variable with cdf (3.1), we write $Y\sim\mathrm{GOLLMax}(\nu,\sigma,\mu)$. The hazard rate function (hrf) of Y is given by $h(y;\mu,\sigma,\nu)=f(y;\mu,\sigma,\nu)/[1-F(y;\mu,\sigma,\nu)]$, where $1-F(y;\mu,\sigma,\nu)$ is survival function.

The GOLLMax distribution contains as special cases the following well-known distributions. For $\sigma=\nu=1$, we obtain the Maxwell (Max) distribution. For $\nu=1$, we have the exponentiated Maxwell (EMax) distribution. For $\sigma=1$, we obtain the odd log-logistic Maxwell (OLLMax) distribution.

The GOLLMax distribution can be simulated by inverting (3.1). The quantile function (qf) of Y, say

$$y=Q_{\mathrm{Max}}\left(\left[\frac{\left(\frac{u}{1-u}\right)^{\frac{1}{\nu}}}{1+\left(\frac{u}{1-u}\right)^{\frac{1}{\nu}}}\right]^{\frac{1}{\sigma}}\right), \tag{3.3}$$

where $Q_{\mathrm{Max}}(u)=G^{-1}(\mu;u)$ is the qf of the Maxwell distribution. The histograms from three simulated data sets form (3.3) and the plots of the exact GOLLMax densities for some parameter values are displayed in Figure 3. These histograms are constructed based on 2,000 generated values. We note that this distribution can fit data with modal and slight bimodality shapes as well as positive and negative skewness.

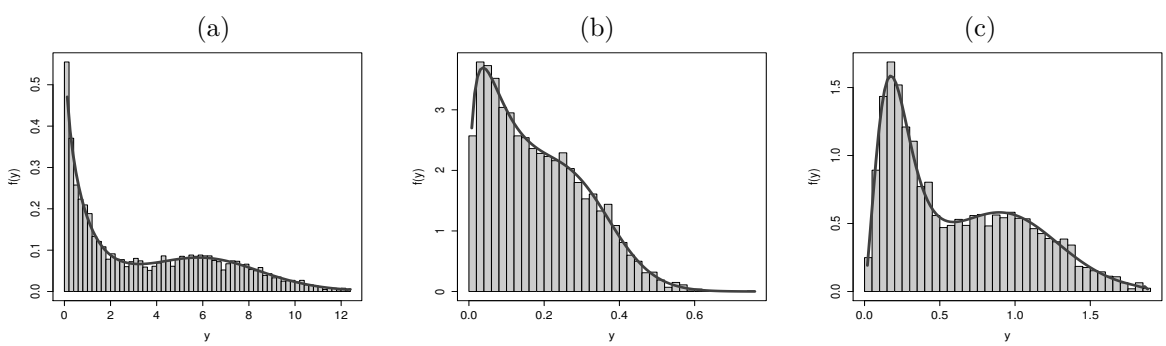

Figure 3.3. Plots of the histograms for simulated values and exact density function (4.2). (a) For $\mu=2.0$, $\sigma=2.0$ and $\nu=0.15$. (b) For $\mu=0.15$, $\sigma=1.0$ and $\nu=0.45$. (c) For $\mu=0.30$, $\sigma=5.0$ and $\nu=0.15$.

3.3 The GOLLMax semiparametric regression model

In many practical applications, the response variable Y can be influenced by explanatory variables such as gender, age, cholesterol level, blood pressure and many others. Let $Y_1,\ldots,Y_n$ are random variable independent and each Y_i has a density function (3.2) and a vector $\mathbf{x}=(x_1,\ldots,x_p)^T$ of exploratory variables, (for $i=1,\ldots,n$). We can consider a regression structure given by

$$g(\mu_i)=\mathbf{x}_i^T\boldsymbol{\beta}, \tag{3.4}$$

where $g : [0, \infty) \to R$ are known one-to-one link functions continuously twice differentiables. The usual systematic component for the scala parameter is $g(\mu_i) = \log(\mu_i)$, then μ_i can be obtained by inverting $g(\mu_i)$, as $\mu = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$, for $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$ is a vector of known explanatory variables for the ith observation, and $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{1p})^T$, are parameter vectors of dimension p. Then, $g(\boldsymbol{\mu}) = \exp(\mathbf{X}\boldsymbol{\beta})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is a specified $n \times p$ matrix of full column rank $p < n$.

Consider a sample $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ of n independent observations. Conventional likelihood estimation techniques for uncensored data can be applied here. The total log-likelihood function for the vector of parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \sigma, \nu)^T$ from model (3.4) takes the form

$$\begin{aligned} l(\boldsymbol{\psi}) &= n \log\left(\frac{4}{\sqrt{\pi}}\right) + \sum_{i=1}^{n} \log(\nu) + \sum_{i=1}^{n} \log(\sigma) + \sum_{i=1}^{n} \log\left(\frac{x_i^2}{\mu_i^3}\right) - \sum_{i=1}^{n} \left(\frac{x_i^2}{\mu_i^2}\right) + \\ & (\sigma\,\nu - 1) \sum_{i=1}^{n} \log\left[\gamma_1(3/2, x_i^2/\mu_i^2)\right] + (\nu - 1) \sum_{i=1}^{n} \log\left[1 - \gamma_1(3/2, x_i^2/\mu_i^2)\right] - \\ & 2 \sum_{i=1}^{n} \log\left\{\gamma_1^{\sigma\,\nu}(3/2, x_i^2/\mu_i^2) + \left[1 - \gamma_1^{\sigma}(3/2, x_i^2/\mu_i^2)\right]^{\nu}\right\}. \end{aligned} \tag{3.5}$$

For censored data, y_i denotes the observed time for the ith subject, i.e, $y_i = \min\{Y_i, C_i\}$, Y_i is the lifetime for the ith individual and C_i is the censoring time for the ith individual. With this assumption we have, that the contribution of an individual that failed at y_i to the likelihood function is given by

$$\begin{aligned} l(\boldsymbol{\psi}) &= r \log\left(\frac{4\,\sigma\,\nu}{\sqrt{\pi}}\right) + \sum_{i \in F} \log\left(\frac{y_i^2}{\mu_i^3}\right) - \sum_{i \in F} \left(\frac{y_i}{\mu_i}\right)^2 + \\ & (\sigma\,\nu - 1) \sum_{i \in F} \log\left[\gamma_1(3/2, y_i^2/\mu_i^2\right] + (\nu - 1) \sum_{i \in F} \log\left[1 - \gamma_1^{\lambda}(3/2, y_i^2/\mu_i^2)\right] - \\ & 2 \sum_{i \in F} \log\left\{\gamma_1^{\sigma\,\nu}(3/2, y_i^2/\mu_i^2) + \left[1 - \gamma_1^{\sigma}(3/2, y_i^2/\mu_i^2)\right]^{\nu}\right\} \\ & + \sum_{i \in C} \log\left\{1 - \frac{\gamma_1^{\sigma\nu}(3/2, y_i^2/\mu_i^2)}{\gamma_1^{\sigma\nu}(3/2, y_i^2/\mu_i^2) + [1 - \gamma_1^{\sigma}(3/2, y_i^2/\mu_i^2)]^{\nu}}\right\}, \end{aligned} \tag{3.6}$$

where r is the number of uncensored observations (failures), F and C denote, respectively, that the set of individuals is a lifetime or a censoring time.

In some cases we can have covariates that have a non-linear relationship whit response variable, so to capture the non-linear effects of these covariates, it is necessary to adopt non-linear functions.

Let $\mathbf{x}_{2i}^{\top} = (x_{2i1}, \ldots, x_{2ip})$ be the vector of covariates that has a nonlinear form with the response variable, we can define semi-parametric structures using appropriate link functions as

$$\mu_i = \exp\left(\mathbf{x}_{1i}^{\top}\boldsymbol{\beta} + \sum_{j=1}^{J} h_j(\mathbf{x}_{j2i})\right) \tag{3.7}$$

where $h_j(\cdot)$ are smooth functions of the covariates $\mathbf{x}_{2i}$ for $j = 1, \ldots, J$ and $i = 1, \ldots, n$. In this paper, the approximation of $h_j(\cdot)$ is by cubic spline. In the `gamlss` package of the `R` software, such smoothing functions are expressed as random effects, i.e. $h_j(\cdot) = Z_j \boldsymbol{\upsilon}_j$, where Z_j is the $(n \times q_j)$ known basis design matrix and $\boldsymbol{\upsilon}_j$ is the q_j-dimensional unknown vector of parameters.

For the semiparametric model (3.7), the fixed and random effects $\boldsymbol{\psi} = (\boldsymbol{\beta}, \sigma, \nu)$ and $\boldsymbol{\varphi}$, respectively, are estimated by maximizing the penalized log-likelihood function

$$l_p(\boldsymbol{\omega}) = l(\boldsymbol{\psi}) - \frac{1}{2} \sum_{j=1}^{J} \lambda_j \boldsymbol{\varphi}_j^T \mathbf{P}_j \boldsymbol{\varphi}_j, \tag{3.8}$$

for $\boldsymbol{\omega} = (\boldsymbol{\beta}, \sigma, \nu, \boldsymbol{\varphi}, \boldsymbol{\lambda})$, where $l(\boldsymbol{\psi})$ can be (3.5) or (3.6), λ_j is the unknown smoothing parameter and P_j is a symmetric matrix that may depend on a vector of smoothing parameters and chosen number of knots. The solution of (3.8) corresponds to the smoothing cubic spline with equidistant knots for distinct x-variable values, more details see (Green and Silverman, 1993, Ruppert, *et al.* 2003 and Wood, 2017). Another measure of interest are the effective degrees of freedom, df^*, relative to the non-parametric component. The smoothing parameters can be fixed or estimated from the data. Some methods are proposed in the literature, for example, by the generalized cross-validation method (see , Wood, 2006). However, in this work, due to the direct relationship between df^* and λ_j, to penalize overfitting the Akaike information criterion AIC (Akaike, 1983) is used.

The numerical maximization of the (3.8) can be performed in the `gamlss` and `gamlss.cens` packages in `R`. We use the maximization by `RS` algorithm described by Stasinopoulos and Rigby (2007) and Stasinopoulos et al. (2017). The `cs()` function is used to assign the arguments to make the adjustment via `gamlss`. Thus, in Section 3.5, the `cs(·)` function in regression structures is denoted by $cs(x_{ji}, df^*)$ where x_{ji} is the j-th covariate considering the additive term and df^* are the degrees of freedom related to the additive term. The effective degree of freedom for structure of the regression in μ considering an explanatory variable x is given by $df_\mu = df^* + 2$ where other two additional degrees of freedom are in relation to the linear terms (see, Voudouris *et al.* (2012)). Finally, we have that the total freedom degree of the adjusted model, represented by df, collectively considers the additive terms represented by the $h_j(\cdot)$ functions and the parametric terms, i.e, $df = df_\mu + df_\sigma + df_\nu$, are the degrees of freedom used to model μ, σ and ν, respectively.

3.3.1 Choosing the best model

For selection of the appropriate distribution, we use the global deviance (GD), $GD = -2l_p(\hat{\boldsymbol{\omega}})$, $l_p(\hat{\boldsymbol{\omega}})$ is the penalized log-likelihood function and the generalized Akaike information criterion (GAIC) defined by $GAIC(k) = GD + k \times df$, where df is the total degrees of freedom of the adjusted model and k is the penalty for each degree of freedom used. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are special cases of the GAIC(k) measure when $k = 2$ and $k = \log(n)$, respectively. We consider the GD, AIC and BIC measures to select the best models.

3.3.2 Influence and residual analysis

The measure to evaluate the influence of an observation (Cook and Weisberg, 1982), called the log-likelihood distance, is the difference between $\hat{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\omega}}_{(i)}$ on the log-likelihood scale, namely

$$LD_i = 2\Big[l_p(\hat{\boldsymbol{\omega}}) - l_p(\hat{\boldsymbol{\omega}}_{(i)})\Big], \tag{3.9}$$

where $l(\hat{\boldsymbol{\omega}})$ is the maximized log-likelihood for the full sample and $l(\hat{\boldsymbol{\omega}}_{(i)})$ is the maximized log-likelihood for the sample excluding the ith observation.

Another important step in the analysis of a fitted model is to check possible deviations from the model assumptions. In this context, we consider the quantile residuals (Dunn and Smyth, 1996) for the GOLLMax semiparametric regression model have the form

$$\widehat{rq}_i = \Phi^{-1}\left\{\frac{\gamma_1^{\hat{\nu}\,\hat{\sigma}}(3/2, y^2/\hat{\mu}_i^2)}{\gamma_1^{\hat{\nu}\hat{\sigma}}(3/2, y^2/\hat{\mu}_i^2) + \left[1 - \gamma_1^{\hat{\sigma}}(3/2, y^2/\hat{\mu}_i^2)\right]^{\hat{\nu}}}\right\}, \tag{3.10}$$

where $\Phi^{-1}(\cdot)$ is the standard normal qf.

We built envelopes to enable better interpretation of the probability normal plot of the residuals. These envelopes are simulated confidence bands described by Atkinson (1985) that contain the residuals, such that if the model is well-fitted, the majority of points will be randomly distributed within these bands.

3.4 Simulation study

In this section, we examine the performance of the GOLLMax semiparametric regression model by means of a Monte Carlo simulation study under two scenarios. Various simulations are conducted for different sample sizes ($n = 80, 200, 450$) for censored and uncensored data using the `R` software with `gamlss` packages by `RS` method.

In this study we present and compare the results, adjusting the proposed GOLLMax model (parametric and semiparametric). The following two scenarios with regression structure (3.4) are presented:

- The first scenario, we consider the following structure for regression model with parameter $\mu_i = \exp\{h_1(x_{1i}) + \beta_{21}x_{2i} + \beta_{31}x_{3i}\}$, where $h_1(x_{1i}) = \sin(x_{1i})$, $\sigma = \exp\{\beta_{02}\}$, $\nu = \exp\{\beta_{03}\}$, with the following sub-scenarios: (a) without censored values and (b) with a percentage of censored values approximately 25%. The functional shape of the $h_1(x_{1i})$ is presented in Figure 3.4a and b, for sub-scenarios (a) and (b), respectively.

- The second scenario, we consider the following structure for regression model with parameter $\mu_i = \exp\{h_1(x_{i1}) + \beta_{21}x_{2i} + \beta_{31}x_{3i}\}$, where $h_1(x_{1i}) = 0.45[\sin(\pi\, x_{1i})]$, $\sigma = \exp\{\beta_{02}\}$, $\nu = \exp\{\beta_{03}\}$, with the following sub-scenarios: (c) without censored values and (d) with a percentage of censored values approximately 25%. The functional shape of the $h_1(x_{1i})$ is presented in Figure 3.4c and d, for sub-scenarios (c) and (d), respectively.

The associated coefficients for the two scenarios are: $\beta_{21} = 0.20$, $\beta_{31} = -0.35$, $\beta_{02} = 0.25$ and $\beta_{03} = 0.40$. We assume that the explanatory variables for the two scenarios are $x_{1i} \sim \text{Uniform}(0, 2.5)$, $x_{2i} \sim \text{Normal}(5, 0.5)$ and $x_{3i} \sim \text{Uniform}(0, 1)$, respectively, for $i = 1, \ldots, n$. For all scenarios it was considered $df^* = 3$ which is the default value of the `cs(·)` function.

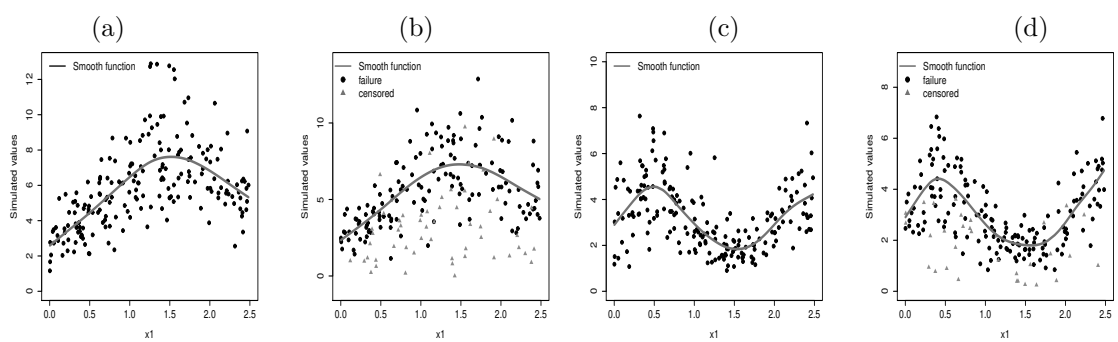

Figure 3.4. Plots of the simulation values for $n = 200$. (a) Nonlinear effect for sub-scenario a. (b) Nonlinear effect for sub-scenario b. (c) Nonlinear effect for sub-scenario c. (d) Nonlinear effect for sub-scenario d.

We simulate the GOLLMax semiparametric regression model under the algorithm:

1. Generate the variables x_{1i}, x_{2i} and x_{3i}.

2. For uncensored values of the sub-scenarios (a) and (c):
   - Generate the values y_i ∼GOLLMax(μ_i, σ, ν) from the structure regression (3.4).

3. For censored values of the sub-scenarios (b) and (d):
   - Generate the values y_i^* ∼GOLLMax(μ_i, σ, ν) from the structure regression (3.4).
   - Generate c_i ∼Uniform$(0, \zeta)$, where ζ denotes the proportion of censored observations.
   - Set $y_i = \min(y_i^*, c_i)$.
   - Define a vector δ of dimension n which receives one if $(y_i^* \leq c_i)$ and zero otherwise.

The samples can be easily generated in R using the code rGOLLMax(n, μ, σ, ν) as shown in Appendix A.

For each of the $1,000$ simulations, the average estimates (AEs), biases and MSEs are calculated. The results are reported in Tables (3.1-3.4) for the parametric and semiparametric models. Based on the simulation results presented, our interest is in verifying how much the inclusion of an additive term affects in the estimations of the other fixed parameters. For semiparametric model, we verify that the MSEs of the MLEs of β_{21}, β_{31}, β_{20} and β_{30} for sub-scenario a, b, c and d decay toward zero when the sample size n increases as far as the case with censored and uncensored values, as usually expected under first-order asymptotic theory. The mean estimates of the parameters tend to be closer to the true parameter values when n increases. However, for the parametric model, such measures do not exhibit the same behavior.

Table 3.1. The AEs, biases and MSEs for the GOLLMax (parametric and semiparametric) regression models based on 1,000 simulations for sub-scenario a.

	Semiparametric				Parametric			
n	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.199	-0.001	0.004	β_{21}	0.115	-0.085	0.012
	β_{31}	-0.352	-0.002	0.012	β_{31}	-0.388	-0.038	0.015
80	β_{02}	0.239	-0.011	0.007	β_{02}	0.300	0.050	0.113
	β_{03}	0.476	0.076	0.014	β_{03}	0.134	-0.266	0.113
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.198	-0.002	0.001	β_{21}	0.214	0.014	0.002
	β_{31}	-0.348	0.002	0.004	β_{31}	-0.317	0.033	0.006
200	β_{02}	0.221	-0.029	0.005	β_{02}	0.250	0.000	0.040
	β_{03}	0.440	0.040	0.006	β_{03}	0.171	-0.229	0.068
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.201	0.001	0.001	β_{21}	0.207	0.007	0.001
	β_{31}	-0.350	0.000	0.002	β_{31}	-0.359	-0.009	0.002
450	β_{02}	0.216	-0.034	0.003	β_{02}	0.258	0.008	0.018
	β_{03}	0.430	0.030	0.003	β_{03}	0.180	-0.220	0.055

In relation to the behavior of the nonlinear effects in the simulations (sub-scenarios a, b, c and d), in the Figures 3.5 displays the generated and fitted effects for the parametric and semiparametric models. We also present in this figure the box-plots of the GD, AIC and BIC statistics obtained in 1,000 simulations for both models. We can note that the nonlinear effects are very close to the true shape as shown in the Figure 3.4, when the sample size increases. Further, we can conclude that the semiparametric model presents the lowest values of GD, AIC and BIC statistics, indicating that it is the most suitable model for simulated data in the presence of non-linear effects.

In Table 3.5 we present a study the actual degrees of freedom of adjustment by cubic spline, considering the same scenarios presented above. In this study it is found that when we have behaviors like point cloud or slightly nonlinear as sub scenarios a and b the default $df^* = 3$ gives good results

Table 3.2. The AEs, biases and MSEs for the GOLLMax (parametric and semiparametric) regression models based on 1,000 simulations for sub-scenario b.

	Semiparametric				Parametric			
n	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.197	-0.003	0.006	β_{21}	0.112	-0.088	0.015
	β_{31}	-0.355	-0.005	0.004	β_{31}	-0.389	-0.039	0.019
80	β_{20}	0.232	-0.018	0.003	β_{02}	0.168	-0.082	0.036
	β_{03}	0.488	0.088	0.016	β_{03}	0.222	-0.178	0.049
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.199	-0.001	0.002	β_{21}	0.216	0.016	0.002
	β_{31}	-0.345	0.005	0.006	β_{31}	-0.318	0.032	0.008
200	β_{02}	0.218	-0.032	0.002	β_{02}	0.155	-0.095	0.019
	β_{03}	0.447	0.047	0.005	β_{03}	0.236	-0.164	0.033
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.200	0.000	0.001	β_{21}	0.205	0.005	0.001
	β_{31}	-0.352	0.002	0.002	β_{31}	-0.361	-0.011	0.003
450	β_{02}	0.211	-0.039	0.002	β_{02}	0.160	-0.090	0.013
	β_{03}	0.434	0.034	0.002	β_{03}	0.245	-0.155	0.027

Table 3.3. The AEs, biases and MSEs for the GOLLMax (parametric and semiparametric) regression models based on 1,000 simulations for sub-scenario c.

	Semiparametric				Parametric			
n	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.186	-0.014	0.004	β_{21}	0.240	0.040	0.006
	β_{31}	-0.366	-0.016	0.012	β_{31}	-0.301	0.049	0.017
80	β_{20}	0.233	-0.017	0.007	β_{02}	0.553	0.303	0.356
	β_{03}	0.450	0.050	0.010	β_{03}	-0.174	-0.574	0.419
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.197	-0.003	0.001	β_{21}	0.192	-0.008	0.002
	β_{31}	-0.353	- 0.003	0.005	β_{31}	-0.436	-0.086	0.013
200	β_{02}	0.216	-0.034	0.006	β_{02}	0.637	0.387	0.301
	β_{03}	0.416	0.016	0.004	β_{03}	0.294	-0.694	0.531
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.200	0.000	0.001	β_{21}	0.187	-0.013	0.001
	β_{31}	-0.361	-0.011	0.002	β_{31}	-0.393	-0.043	0.004
450	β_{02}	0.212	-0.038	0.004	β_{02}	0.476	0.226	0.108
	β_{03}	0.406	0.006	0.002	β_{03}	-0.172	-0.572	0.347

in smoothing the points. When relationships with sinusoidal forms such as c and d sub-scenarios exist, higher degrees of freedom or the choice of another type of smoothing function may be required.

3.5 Applications

In this section, we provide two applications to real data to prove empirically the flexibility of the GOLLMax semiparametric regression model. The computations are performed using the `gamlss` package in `R` software. In the first application the modeling is performed for an uncensored data. In the second application we consider modeling for censored data.

Table 3.4. The AEs, biases and MSEs for the GOLLMax (parametric and semiparametric) regression models based on 1,000 simulations for sub-scenario d.

	Semiparametric				Parametric			
n	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
80	β_{21}	0.184	-0.016	0.006	β_{21}	0.242	0.042	0.008
	β_{31}	-0.370	-0.020	0.014	β_{31}	-0.316	0.034	0.019
	β_{20}	0.219	-0.031	0.004	β_{02}	0.377	0.127	0.198
	β_{03}	0.457	0.057	0.010	β_{03}	-0.072	-0.472	0.298
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
200	β_{21}	0.197	-0.003	0.002	β_{21}	0.191	-0.009	0.002
	β_{31}	-0.354	0.004	0.005	β_{31}	-0.437	-0.087	0.015
	β_{02}	0.205	-0.045	0.003	β_{02}	0.481	0.231	0.183
	β_{03}	0.420	0.020	0.003	β_{03}	-0.194	0.594	0.401
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
450	β_{21}	0.197	-0.003	0.001	β_{21}	0.186	-0.014	0.001
	β_{31}	-0.363	-0.013	0.002	β_{31}	-0.394	-0.044	0.005
	β_{02}	0.200	-0.050	0.003	β_{02}	0.383	0.133	0.068
	β_{03}	0.408	0.008	0.001	β_{03}	-0.011	-0.510	0.280

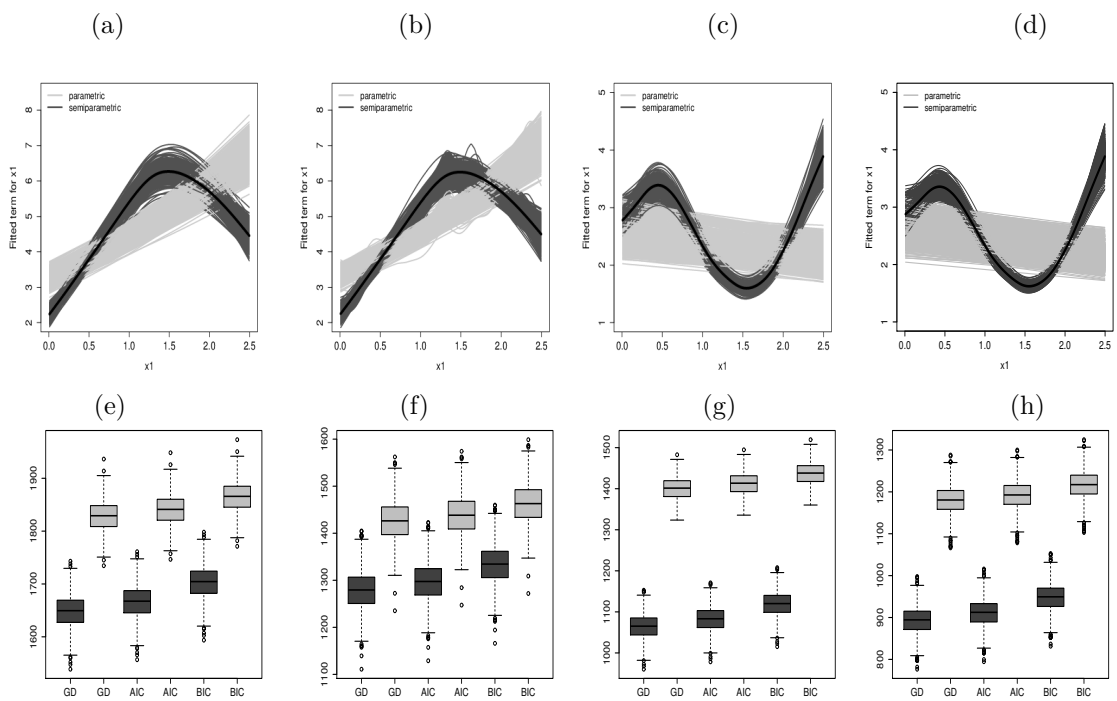

Figure 3.5. The fitted GOLLMaxSemi regression model under scenarios a, b, c and d for $n = 450$. (a) Fitted of the effect of x_1 in μ parameter for scenario 1. (b) Fitted of the effect of x_1 in μ parameter for scenario 2. (c) Fitted of the effect of x_1 in μ parameter for scenario 3. (d) Fitted of the effect of x_1 in μ parameter for scenario 4. (e) Goodness-of-fit statistics measures for scenario 1. (f) Goodness-of-fit statistics measures for scenario 2. (g) Goodness-of-fit statistics measures for scenario 3. (h) Goodness-of-fit statistics measures for scenario 4.

3.5.1 Application 1: uncensored data

In the first application, we consider the relationship of two variables that are part of the data set analyzed by Kelly *et. al*, (2012). In this study the authors propose the development of a tool to predict,

Table 3.5. The λ estimates and average AIC measurement for GOLLMax semiparametric regression models based on 1,000 simulations for sub-scenarios a, b, c and d.

	sub-scenario a		sub-scenario b		sub-scenario c		sub-scenario d	
df^*	λ	AIC	λ	AIC	λ	AIC	λ	AIC
0	-	824.30	-	647.63	-	649.23	-	545.17
1	0.191	750.27	0.197	591.95	0.183	569.28	0.195	481.12
2	0.038	740.15	0.038	579.63	0.037	515.49	0.038	434.69
3	**0.012**	**739.95**	**0.012**	**579.17**	0.011	492.58	0.012	412.22
4	0.004	740.47	0.004	579.75	0.004	486.22	0.002	403.32
5	0.002	741.10	0.002	580.46	**0.002**	**484.91**	**0.001**	**403.26**
6	0.001	741.73	0.001	581.18	0.001	484.94	0.001	403.68
7	0.001	742.37	0.001	581.90	0.001	485.36	0.000	404.25
8	0.000	742.99	0.000	582.61	0.000	485.90	0.000	404.79
9	0.000	743.62	0.000	583.32	0.000	486.48	0.000	405.37
10	0.000	744.22	0.000	584.02	0.000	487.07	0.000	405.52

in the preoperative period, the need for a patient to attend an extended care service after orthotopic liver transplantation. As a way of illustrating the applicability of the model we propose, the variables considered are: (y = BMI) and (x_1 = age), respectively, of transplanted patients.

First, we present an exploratory and marginal analysis of the BMI data. The data set with $n = 777$ observations. Table 3.6 provides the mean, median, standard deviation, skewness, kurtosis, minimum and maximum measures. In this table, stands out the value in relation to the skewness these suggest positively skewed distributions.

Table 3.6. Descriptive Statistics for the BMI data.

Mean	Median	SD	Skewness	Kurtosis	Min.	Max.
28.648	28.0	5.967	0.679	0.734	15.0	55.0

In this way, we consider in this analysis the proposed GOLLMax model and the particular OLLMax case. We also consider for comparison the Weibull and normal models, which are frequently used in data analysis. Figure 3.6a the adjusted curves of the density functions considered under the marginal analysis of the response variable y are presented. As previously mentioned due to positive asymmetry the normal model visually is not the most suitable. Figure 3.6b shows a box plot of the response variable which indicates the presence of asymmetry and possible outliers. In Figure 3.6c displays the dispersion observed y against x_1 with fitted smooth curve.

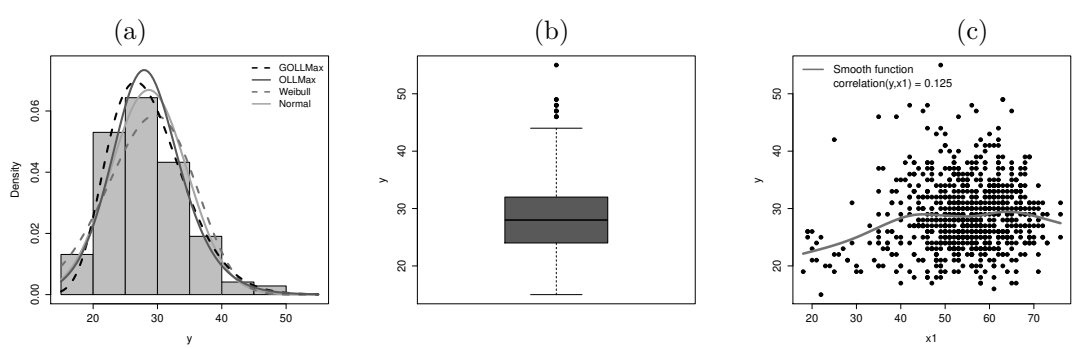

Figure 3.6. Plots for BMI data. (a) Histogram and estimated GOLLMax, OLLMax, Weibull and normal densities. (b) Box-plot for response variable y. (c) Observed y against x_1 with fitted smooth curve.

Further, we present results for the GOLLMax, OLLMax, Weibull and normal semiparametric regression models by considering the following systematic structures as presented in Section 3.3:

$$\textbf{GOLLMax}\left\{\ \mu_i = \exp\{\beta_{10} + cs(x_{1i}, df^*)\},\quad \sigma = \exp\{\beta_{20}\}\quad \text{and}\quad \nu = \exp\{\beta_{30}\};\right.$$
$$\textbf{OLLMax}\left\{\ \mu_i = \exp\{\beta_{10} + cs(x_{1i}, df^*)\}\quad \text{and}\quad \sigma = \exp\{\beta_{20}\};\right.$$
$$\textbf{Weibull}\left\{\ \mu_i = \exp\{\beta_{10} + cs(x_{1i}, df^*)\}\quad \text{and}\quad \sigma = \exp\{\beta_{20}\};\right.$$
$$\textbf{Normal}\left\{\ \mu_i = \beta_{10} + cs(x_{1i}, df^*)\quad \text{and}\quad \sigma = \exp\{\beta_{20}\}.\right.$$

The values of the GD, AIC and BIC statistics and the effective degrees of freedom and estimate for smoothing parameter, λ, of the fitted semiparametric regression models are listed in Table 3.7. Also based on the likelihood ratio test between the GOLLMax and OLLMax models, for null hypotheses $H_0 : \nu = 1$, the obtained test value is ($w = 1.835$) with (p-value $= 0.175$). In this case not reject the null hypotheses and the particular model is the most indicated. By comparing these figures, we conclude that the OLLMax semiparametric regression model outperforms the GOLLMax, Weibull and normal models irrespective of the criteria and then the proposed regression model can be used effectively in the analysis of these data.

Table 3.7. The GD, AIC, BIC, df, df^* and $\hat{\lambda}$ measurements for the GOLLMax, OLLMax, Weibull and Normal semiparametric regression models for the BIM data.

Model	GD	AIC	BIC	df	df^*	$\hat{\lambda}$
GOLLMax	4904.09	4922.09	4963.99	9	5	0.0056
OLLMax	4905.93	4921.93	4959.17	8	5	0.0056
Weibull	5029.35	5045.35	5082.58	8	5	0.0060
Normal	4949.49	4963.48	4996.07	7	4	0.0121

In Figure 3.7 it is shown how the df^* values were chosen. In this way, the λ is estimated according to the choice of df^* that minimizes the AIC measurement.

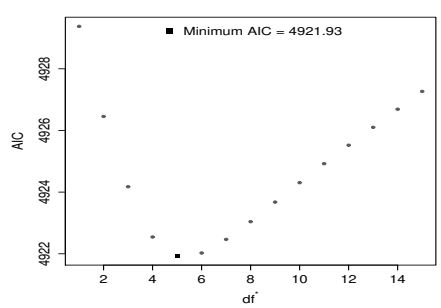

Figure 3.7. Plots of the AIC versus df^* for OLLMax model.

Table 3.8 lists the MLEs, SEs and p-values obtained from the fitted OLLMax semiparametric regression model to the BMI data. This table reveals (at the 5% significance level), considering the proposed systematic component, that the intercept related to the parameter of the term of the nonlinear effects is significant. The coefficients of the nonlinear term is presented, however such values are not interpretable.

Table 3.8. MLEs of the parameters and SEs from the fitted OLLMax semiparametric regression model to the BMI data.

Parameter	Estimate	SE	p-value
β_{10}	3.096	0.038	$<< 0.001$
$cs(x_1, df^* = 5.0)$	0.003	0.007	< 0.001
β_{20}	0.876	0.030	-

We compute case-deletion measures LD_i defined in subsection 3.3.2. The results of such influence measure index plots are displayed in Figure 3.8a. In Figure 3.8b shows the scatter plot with possible influential points identified. They reveal that the observations ♯224, ♯618 and ♯685 are possible

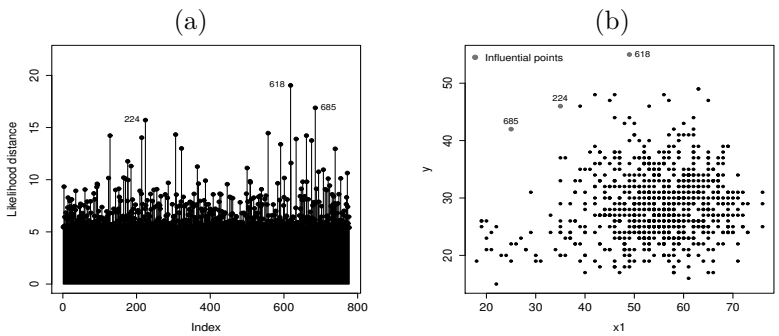

Figure 3.8. Plots for the fitted OLLMax semiparametric regression model to the BMI data. (a) Index plot versus likelihood distance. (b) Observed y against x_1 with influential points.

influential observations. In Table 3.9 gives the relative change (in percentage) of each estimate defined by $RC_{\boldsymbol{\theta}_j} = [(\hat{\boldsymbol{\omega}}_j - \hat{\boldsymbol{\omega}}_j(I))/\hat{\boldsymbol{\omega}}_j]\times 100$, and the corresponding p-value, where $\hat{\boldsymbol{\omega}}_j(I)$ is the MLE of $\boldsymbol{\omega}_j$ after the "set I" of observations being removed. Table 3.9 provides the following sets: $I_1 = \{\sharp 224\}$, $I_2 = \{\sharp 618\}$ and $I_3 = \{\sharp 685\}$.

Table 3.9. MLEs, p-values (in parentheses), and their relative changes [-RC- in %] for the corresponding set.

Dropping	None	Set I_1	Set I_2	Set I_3
β_{10}	3.096	3.090	3.102	3.146
	($<< 0.001$)	($<<0.001$)	($<<0.001$)	(<0.001)
	-	[0.161]	[-0.193]	[-1.614]
$cs(x_1, df^* = 5.0)$	0.003	0.003	0.003	0.002
	(<0.001)	(<0.001)	(<0.001)	(0.003)
	-	[0]	[0]	[33.33]

Based on the in Table 3.9, we note that the MLEs of the parameters of the OLLMax semiparametric regression model are robust to the deletion of influential observations. Moreover, the significance of the estimates of the parameters does not change (at the 5% significance level) after removal of these cases, that is, no changes inferential after removal of observations considered influential in the diagnostics plots. Therefore, the observations are kept in the data set.

In Figure 3.9, we perform the residual analysis by plotting the quantile residuals rq_i (see subsection 3.3.2) against the index of observations for the fitted OLLMax (Figure 3.9a), Weibull (Figure 3.9b) and Normal (Figure 3.9c) regression models. In Figure 3.9a, we note that all the observations are included in the interval $[-3, 3]$ except the cases ♯224, ♯618 and ♯685, and that the residuals appear to

behave randomly. Then, there is no evidence that the model assumptions are inadequate. We can note in Figures 3.9b and 3.9c there are many discrepant points.

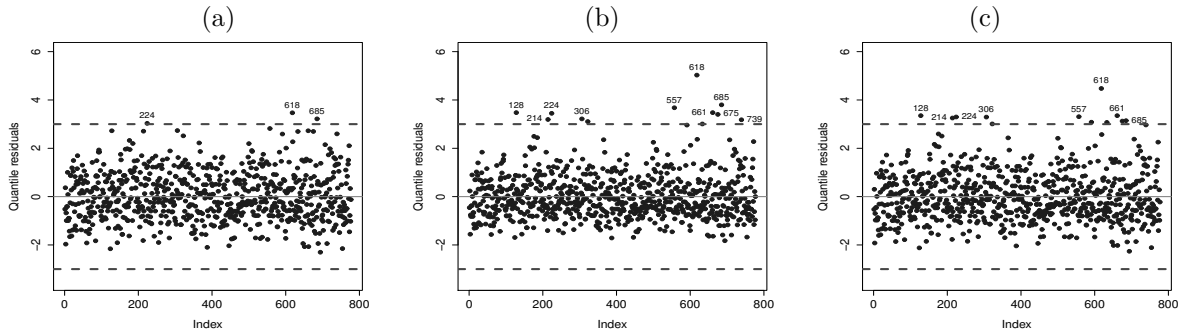

Figure 3.9. The index plot of the quantile residuals with range $[-3, 3]$ for the fitted semiparametric models to the BMI data. (a) OLLMax model. (b) Weibull model . (c) Normal model.

For OLLMax semiparametric regression model is displayed in Figure 3.10 the plots of the quantile residuals and the envelope. This provides indications of the absence of discrepant observations and that the OLLMax model is adequate for this analysis.

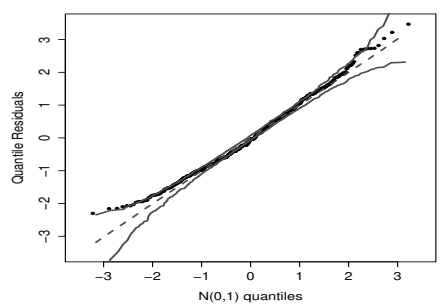

Figure 3.10. Normal probability plot for the quantile residuals with envelope from the fitted OLLMax semiparametric regression model to BMI data.

Figure 3.11a indicates that the BMI increases rapidly until the 40 year of age, and then remains approximately constant from 40 to 60 years of age, declines for age > 60. Figure 3.11b gives six fitted percentile curves q x 100 = (5, 25, 50, 75, 90, 97.5) for y versus the variables x_1. Figure 3.11c exhibits the functional form curve fitted by the OLLMax semiparametric regression model for the BMI data with some conditional densities adjusted to different values of x_1. We finish up that the OLLMax semiparametric regression model can be chosen as the best model.

3.5.2 Application 2: censored data

In this application, we use the monoclonal gammopathy of undetermined significance (MGUS) dataset included in the library `suvival` in `R` software. The plasma cells are responsible for manufacturing immunoglobulins, an important part of the immune defense. At any given time there are estimated to be about 10^6 different immunoglobulins in the circulation at any one time. When a patient has a plasma cell malignancy the distribution will become dominated by a single isotype, the product of the malignant clone, visible as a spike on a serum protein electrophoresis. Monoclonal gammopathy of undertermined

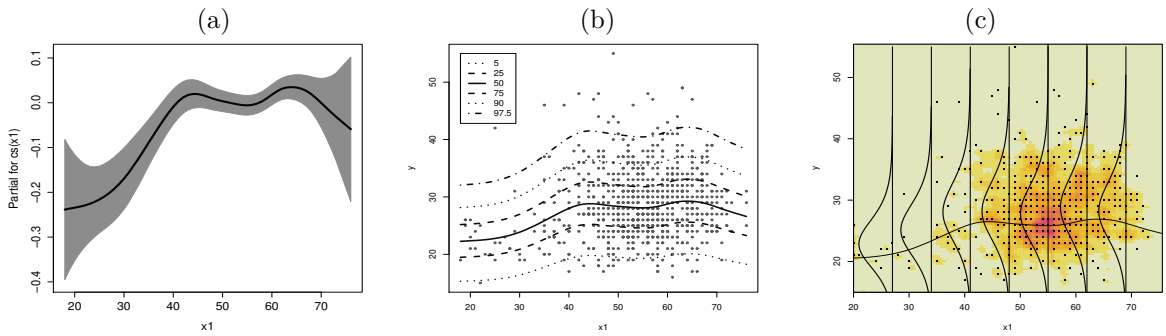

Figure 3.11. The OLLMax semiparametric regression model fitted for the BMI data. (a) Fitted partial effects of the x_1 . (b) Fitted percentile curves for q x 100 = (5, 25, 50, 75, 90, 97.5) against x_1. (c) Smoothed scatterplot diagram which shows how the fitted conditional distribution of the response variable BMI changes for different values of x_1.

significance (MGUS) is the presence of such a spike, but in a patient with no evidence of overt malignancy more details (see Kyle, 1993). The data frame has 241 observations, in this paper we consider the following variables:

- y_i: days from diagnosis to last follow-up;
- x_{i1}: age in years at the detection of MGUS;
- x_{i2}: size of the monoclonal protein spike at diagnosis;
- x_{i3}: sex, a factor with level Male and reference Female,

where $i = 1, \ldots, 240$ because the variable x_{i2} contains an `NA` that was omitted in this analysis.

First, we present an exploratory and marginal analysis of the data. In Figure 3.12a boxplot by x_3 is presented considering failure times, it is visually verified that there is similarity between sex. Figure 3.12b and c displays the dispersion plot between y and covariates x_1 and x_2, respectively.

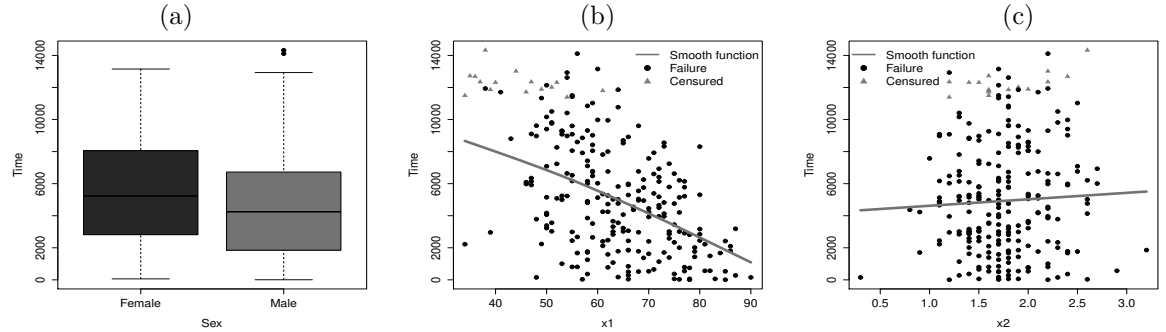

Figure 3.12. Plots for MGUS data. (a) Box plot by x_3. (b) Observed y against x_1 with fitted smooth curve for uncensored values. (c) Observed y against x_2 with fitted smooth curve for uncensored values.

We consider in a preliminary analysis only the times of survival and censoring without covariates. The TTT-plot is displayed in Figure 3.13a, which indicates that the hrf associated with the data set has a increscent shape. Therefore, of the GOLLMax, OLLMax and Weibull distributions can be considered to model these data. In Figure 3.13b displays the estimated survival functions of the GOLLMax,

OLLMax and Weibull distributions and the empirical survival function. The Figure 3.13c displays the estimated hrfs of the GOLLMax, OLLMax and Weibull distributions and the empirical hrf. It turns out that the GOLLMax distribution provides a good fit to the current data for the response variable.

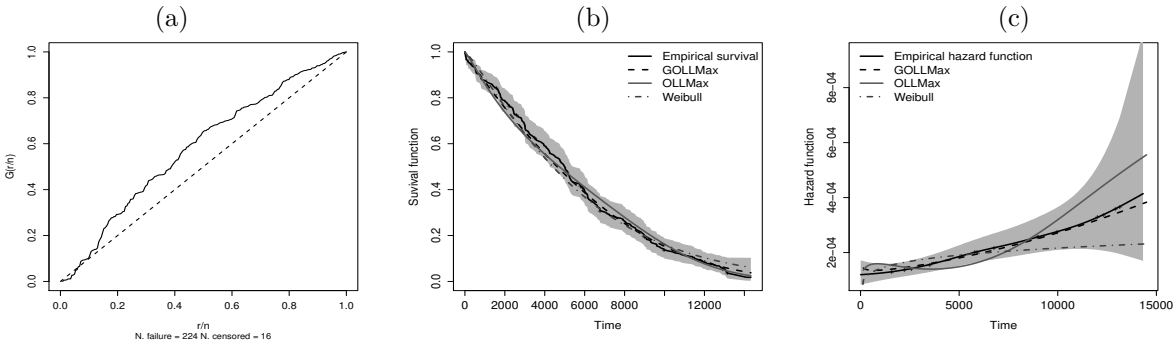

Figure 3.13. Plots for MGUS data. (a) TTT-plot for y. (b) Empirical survival function with the estimated by GOLLMax, OLLMax and Weibull models. (c) Empirical hazard function with the estimated by GOLLMax, OLLMax and Weibull models.

Following with the analysis we consider the three models with regression structure. Unlike application 1 we have in this case two continuous covariates. Thus, to obtain the values of df^* and λ we adopted the following strategy:

i - For obtain df^* and $\hat{\lambda}$ with respect to x_1 we consider x_2 and x_3 fixed with the following structure:$\mu_i = \exp\{\beta_{10} + cs(x_{1i}, df^*) + \beta_{12}x_{2i} + \beta_{13}x_{3i}\}$;

ii - For obtain df^* and $\hat{\lambda}$ with respect to x_2 we consider x_1 and x_3 fixed with the following structure:$\mu_i = \exp\{\beta_{10} + \beta_{11}x_{1i} + cs(x_{2i}, df^*) + \beta_{13}x_{3i}\}$.

Figure 3.14 it is shown how the df^* values were chosen for GOLLMax semiparametric regression model. According to the values presented in Figure 3.14a by regression structure (i) the model for variable x_1 can be considered linear, in agreement with the initial exploratory shown in Figure 3.12b. In Figure 3.14b by regression structure (ii) there is a minimum point indicating the need to use a smoothing function for variable x_2. The same procedure applies for OLLMax and Weibull models, respectively.

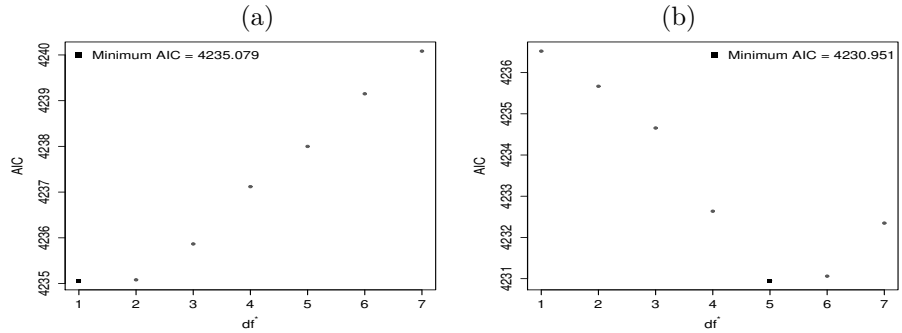

Figure 3.14. Plots of the AIC versus df^* for GOLLMax model for MGUS data. (a) For x_1. (b) For x_2.

Further, we present results for the GOLLMax, OLLMax and Weibull semiparametric regression

models for censored data following systematic structures as presented in Section 3.3:

$$\textbf{GOLLMax}\left\{\ \mu_i = \exp\{\beta_{10} + \beta_{11}x_{1i} + cs(x_{2i}, df^*) + \beta_{13}x_{3i}\},\quad \sigma = \exp\{\beta_{20}\}\quad \text{and}\quad \nu = \exp\{\beta_{30}\};\right.$$

$$\textbf{OLLMax}\left\{\ \mu_i = \exp\{\beta_{10} + \beta_{11}x_{1i} + cs(x_{2i}, df^*) + \beta_{13}x_{3i}\}\quad \text{and}\quad \sigma = \exp\{\beta_{20}\};\right.$$

$$\textbf{Weibull}\left\{\ \mu_i = \exp\{\beta_{10} + \beta_{11}x_{1i} + cs(x_{2i}, df^*) + \beta_{13}x_{3i}\}\quad \text{and}\quad \sigma = \exp\{\beta_{20}\};\right.$$

The values of the GD, AIC and BIC statistics and the effective degrees of freedom and estimate for smoothing parameter, λ, of the fitted semiparametric regression models are listed in Table 3.10. Also based on the likelihood ratio test between the GOLLMax and OLLMax models, for null hypotheses $H_0 : \nu = 1$, the obtained test value is ($w = 7.645$) with (p-value $= 0.005$). In this case reject the null hypotheses and the GOLLMax model is the most indicated. By comparing these figures, we conclude that the GOLLMax semiparametric regression model outperforms the OLLMax and Weibull models irrespective of the criteria. The proposed regression model can be used effectively in the analysis of these data with censoring.

Table 3.10. The GD, AIC, BIC, df, df^* and $\hat{\lambda}$ measurements for the GOLLMax, OLLMax and Weibull semiparametric regression models for the BIM data.

Model	GD	AIC	BIC	df	df^*	$\hat{\lambda}$
GOLLMax	4208.95	4230.95	4269.24	11	5	0.00095
OLLMax	4216.59	4236.59	4271.40	10	5	0.00098
Weibull	4226.56	4248.56	4286.86	11	6	0.00085

Table 3.11 lists the MLEs, SEs and p-values obtained from the fitted GOLLMax semiparametric regression model to the MGUS data. This table reveals (at the 5% significance level), considering the proposed systematic component, that the intercept related to the parameter of the term of the nonlinear effects is significant. The coefficients of the linear term for variable x_1 is significant. This indicates that age has an effect on the life time of patients. Due to the linear decreasing trend, we have that as the age of the patient increases at the time of diagnosis the lifetime decreases. The coefficients of the nonlinear term is presented, however such values are not interpretable. In this set of variables considered, the variable x_3 (sex) also has no significant effect in relation to the lifetime of the patients under study.

Table 3.11. MLEs of the parameters and SEs from the fitted GOLLMax semiparametric regression model to the BMI data.

Parameter	Estimate	SE	p-value
β_{10}	10.899	0.459	$<< 0.001$
β_{11}	-0.037	0.005	$<< 0.001$
$cs(x_2, df^* = 5.0)$	0.099	0.150	0.509
β_{13}	-0.131	0.117	0.265
β_{20}	-0.657	0.058	-
β_{30}	-0.261	0.050	-

We compute case-deletion measures LD_i defined in subsection 3.3.2. The results of such influence measure index plots are displayed in Figure 3.15a. In Figure 3.15b and c shows the scatter plot with possible influential points identified for variables x_1 and x_2, respectively.

They reveal that the observations ♯14 and♯90 are possible influential observations. In Table 3.12 gives the relative change (in percentage) of each estimate defined by $RC_{\boldsymbol{\theta}_j} = [(\hat{\boldsymbol{\omega}}_j - \hat{\boldsymbol{\omega}}_j(I))/\hat{\boldsymbol{\omega}}_j]\times 100$,

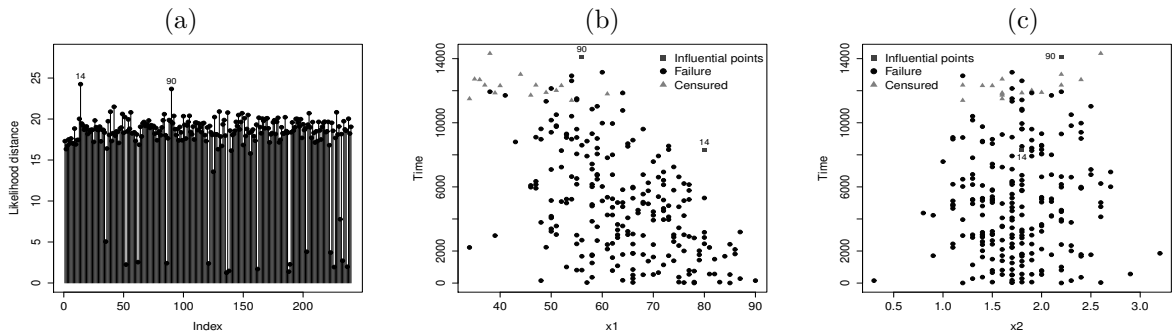

Figure 3.15. Plots for the fitted OLLMax semiparametric regression model to the BMI data. (a) Index plot versus likelihood distance. (b) Observed y against x_1 with influential points. (c) Observed y against x_2 with influential points.

and the corresponding p-value, where $\hat{\boldsymbol{\omega}}_j(I)$ is the MLE of $\boldsymbol{\omega}_j$ after the "set I" of observations being removed. Table 3.9 provides the following sets: $I_1 = \{\sharp 14\}$ and $I_2 = \{\sharp 90\}$.

Table 3.12. MLEs, p-values (in parentheses), and their relative changes [-RC- in %] for the corresponding set.

Dropping	None	Set I_1	Set I_2
	10.899	11.011	10.798
β_{10}	(<< 0.001)	(<<0.001)	(<<0.001)
	-	[-1.027]	[0.917]
	-0.037	-0.039	-0.037
β_{11}	(<< 0.001)	(<<0.001)	(<<0.001)
	-	[-8.108]	[0]
	0.099	0.094	0.106
$cs(x_1, df^* = 5.0)$	(0.509)	(0.520)	(0.497)
	-	[5.050]	[-7.070]
	-0.131	-0.156	-0.145
β_{13}	(0.265)	(0.173)	(0.222)
	-	[-19.847]	[-11.450]

Based on the in Table 3.12, we note that the MLEs of the parameters of the GOLLMax semiparametric regression model are robust to the deletion of influential observations. Moreover, the significance of the estimates of the parameters does not change (at the 5% significance level) after removal of these cases.

In Figure 3.16, we perform the residual analysis by plotting the quantile residuals rq_i (see subsection 3.3.2) against the index of observations for the fitted GOLLMax (Figure 3.16a), OLLMax (Figure 3.16b) and Weibull (Figure 3.16c) semiparametric regression models. In Figure 3.16a, we note that all the observations are included in the interval $[-3, 3]$ except the cases $\sharp 79$ and $\sharp 135$, and that the residuals appear to behave randomly. Then, there is no evidence that the model assumptions are inadequate. We can note in Figures 3.16b and 3.16c there are some discrepant points.

Figure 3.17 shows the plots of the quantile residuals and the simulated envelope for GOLLMax semiparametric regression model. This provides indications of the absence of discrepant observations and that the GOLLMax model is adequate for this analysis.

The partial effects for the covariates in relation to the systematic structures are displayed in Figures 3.18. In Figure 3.18a, we present the effect of the term linear for variable x_1, as we have already interpreted before, when the age of the patient is elevated at the moment of diagnosis the lifetime

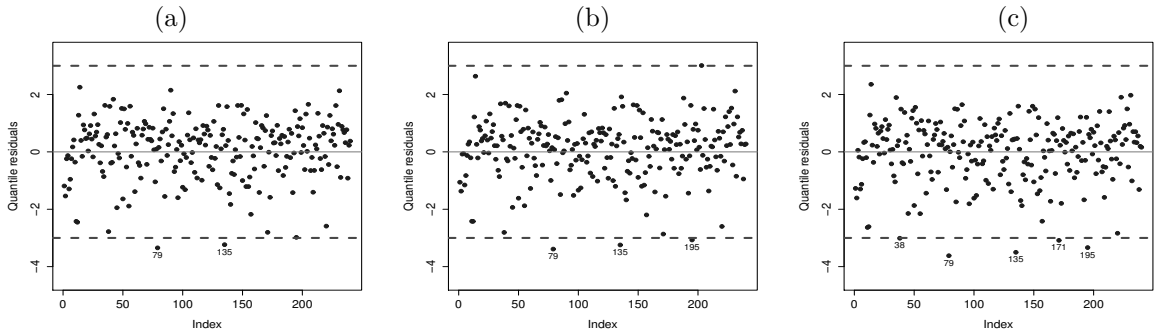

Figure 3.16. Plots of the index versus quantile residuals for MGUS data. (a) GOLLMax model. (b) OLLMax model . (c) Weibull model.

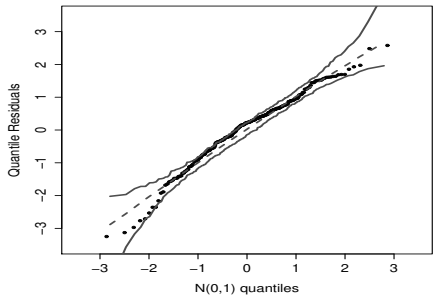

Figure 3.17. Normal probability plot for the quantile residuals with envelope from the fitted GOLLMax semiparametric regression model to MGUS data.

decreases. Figure 3.18b indicates that for monoclonal protein spike measurements between approximately 0.5 and 1 there is a growing linear trend in lifetime. For values between 1 and 2.6 the lifetime it remains constant and from 2.6 there is a linear decreasing trend in the lifetime.

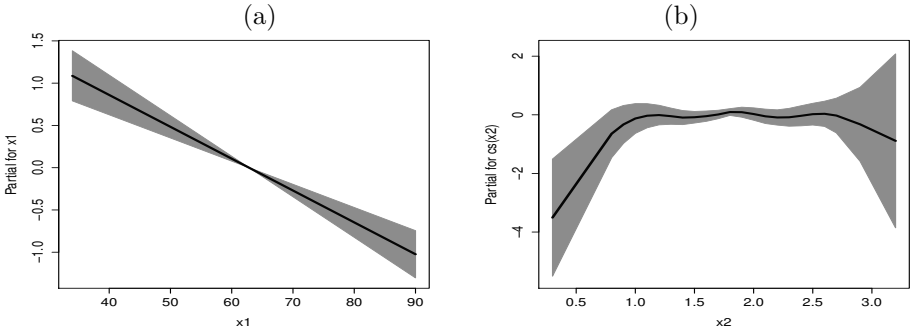

Figure 3.18. The GOLLMax semiparametric regression model fitted for the MGUS data. (a) Fitted partial effects of the x_1. (b) Fitted partial effects of the x_2.

3.6 Concluding Remarks

The *generalized odd log-logistic Maxwell* semiparametric regression model provides a flexible regression model for a dependent real outcome. It's defined a GOLLMax regression model with systematic

components based on the new distribution, which is very suitable for modeling censored and uncensored data. Procedures for fitting the semiparametric GOLLMax regression model and for model diagnostics are included in the `gamlss` package and available from the authors. Two real data sets are used to illustrate the importance of the semiparametric GOLLMax regression model, considering censored and uncensored data.

References

Atkinson, A.C. (1985). *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis*. Clarendon Press, Oxford.

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

Kelly, D. M., Bennett, R., Brown, N., McCoy, J., Boerner, D., Yu, C., ... and Kattan, M. W. (2012). Predicting the discharge status after liver transplantation at a single center: a new approach for a new era. **Liver Transplantation**, **18**, 796-802.

Kyle, R. A. (1993). ´´Benign´´ monoclonal gammopathy-after 20 to 35 years of follow-up. *In Mayo Clinic Proceedings*, **68**, 26-36

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric regression*. Cambridge university press.

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

Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., Voudouris V. and De Bastiani, F. (2017), *Flexible Regression and Smoothing: Using GAMLSS in R*, Chapman and Hall/CRC, New York.

Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J. and Stasinopoulos, D.(2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, **39**, 1279-1293.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

4 ZERO ADJUSTED GENERALIZED ODD LOG-LOGISTIC MAXWELL SEMIPARAMETRIC REGRESSION MODEL

Abstract: In various applications, it is common to find data with the presence of bimodality, heteroskedasticity, zero-inflation and covariables with linear and nonlinear effects in relation to the response variable. Therefore, the objective of this chapter is to propose a regression model able to model data in the presence of all these problems. We use the maximum likelihood method to estimate the parameters of the proposed regression model. For different fixed parameters, sample sizes and percentages of zeros, we perform various simulations to assess the behavior of the estimators. Additionally, we develop an analysis of the residuals based on the residual quantile approach to evaluate the assumptions of the proposed model. Finally, the model is illustrated using data from an experiment conducted to assess the soil microbiology in a sugarcane field.
Keywords: Data analysis; Maxwell distribution; Maximum likelihood estimation; Variation of soil microbiological data; Zero adjusted.

4.1 Introduction

Regression analysis is a commonly used statistical technique applied in many scientific fields. The linear regression model with normal distribution is generally used to model data having symmetric distribution. However, various phenomena cannot always be modeled with the normal distribution, be it for the lack of symmetry, the existence of bimodality or the presence of atypical values.

In past decades, when the phenomenon of interest did not satisfy the assumption of normality of the response variable, some type of transformation was applied at least to obtain symmetric behavior of the data. However, recently it has become more attractive to propose new regression models to model different types of data.

For example, in the area of microbiology, various phenomena are observed where the data are not normally distributed or have other problems, such as asymmetry, non-constant variance (heteroskedasticity) or bimodality.

Regression models have permitted understanding many aspects of the interaction of soil with crops and climate factors. In particular, the use of molecular tools has enabled identifying components of previously unknown microbial communities, increasing the range of information regarding soil quality (Lambais *et al.* 2005 and Andreote *et al.* 2017). However, many of the datasets generated by application of molecular techniques, such as terminal restriction fragment length polymorphism (`T-RFLP`), have peculiar distributions. Each dataset obtained is a reflection of the heterogeneity of the soil where the sample was collected. In Brazil, molecular analytic techniques have attracted particular attention in the sugar-alcohol sector, aiming to reduce the traditional application of mineral fertilizers to grow sugarcane, and instead use microbiota from the soil with application of organic matter. One of these alternatives (Gurdeep and Reddy, 2015, Estrada-Bonilla *et al.* 2017 and Soltangheisi *et al.* 2019) is to use microorganisms that can enhance the availability to plants of nutrients, especially phosphorus. However, the results obtained by experiments in this respect are hard to interpret due to limitations in the use of statistical techniques.

An experiment was recently conducted by the Department of Soil Science of the Luiz de Queiroz College of Agriculture of the University of São Paulo employing a molecular approach in field conditions. The objective of this study is to discover which factors positively or negatively influence the biological response of the soil to the different treatments applied in that experiment, by applying a regression model

to that set of microbiological data able to reflect as closely as possible the productive gain observed in the field.

In this study, the response variable (Y) is the terminal restriction fragment length polymorphism (`T-RFLP`) and the covariables or explanatory variables are, for example, the abundance of the gene that encodes the `16S RNAr` subunit (for bacteria). The region `ITS` corresponds to the group of total fungi in the soil, and the gene *phoD*, which encodes the enzyme alkaline phosphatase. We also consider as covariables the treatments applied in the experiment. The description of these variables is presented in the section 4.6.

According to the descriptive analysis of the response variable (y=`T-RFLP`), we have the following observations:

- Possible presence of bimodality in the dataset according to the distribution of Y (Figure 4.1a).
- High percentage of zeros in the response variable (Figures 4.1b and 4.1c).
- Presence of heterogeneity and nonlinear behavior of the covariables (x_1 = `16S RNAr` , x_2 = `ITS` and x_3 = *phoD*) and the response variable Y (Figures, 4.2a, 4.2b and 4.2c).

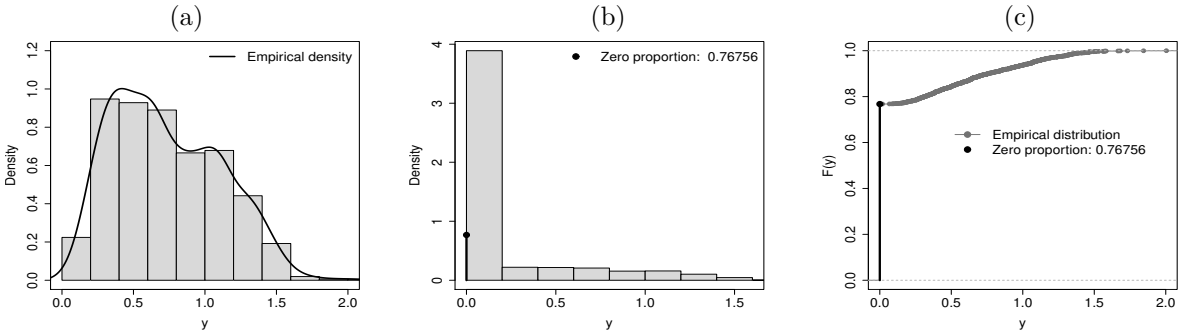

Figure 4.1. Plots of the microbiological data. (a) Histogram with empirical density for y without the zeros. (b) Histogram as for y with the zeros. (c) Empirical distribution for y with zeros .

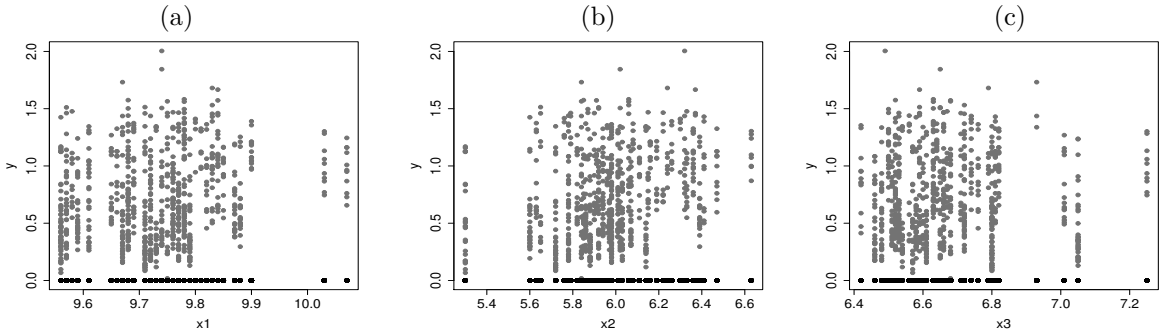

Figure 4.2. Plots of the dispersion for response variable against covariables. (a) Dispersion of y against x_1. (b) Dispersion of y against x_2. (c) Dispersion of y against x_3.

Based on the characteristics observed in the set of microbiological data (Figures 4.1 and 4.2), we propose the following new models:

- First we propose a new distribution, called *generalized odd log-logistic Maxwell* (GOLLMax), based on the Maxwell distribution.
- Based on the GOLLMax distribution, we propose the *zero adjusted generalized odd log-logistic Maxwell* model (ZAGOLLMax) to model the excess of zeros.
- Then we propose regression models based on the ZAGOLLMax model considering two systematic components, where one of these components models the presence of heterogeneity in the data.
- Finally, we propose a semiparametric zero-inflated regression model with the objective of modeling the nonlinear effect, based on the ZAGOLLMax distribution.

After fitting the model, it is important to check the model assumptions and conduct robustness studies to detect possible influential or extreme observations that can cause distortions on the results of the analysis. In this chapter, we discuss the diagnostic influence of the ith observation on the parameter estimates by removing it from the analysis.We propose diagnostic measures based on case-deletion for the ZAGOLLMax semiparametric regression model in order to determine which subject might be influential in the analysis.

The assessment of the fitted model is an important part of data analysis, particularly in regression models, and residual analysis is a helpful tool to validate the fitted model. Examination of residuals can be used, for instance, to detect the presence of outlying observations, the absence of components in the systematic part of the model and departures from the error and variance assumptions. In this chapter, we proposed a residual quantile for the ZAGOLLMax semiparametric regression model whose empirical distribution is close to normality.

The plan of the following sections of the chapter are as follows. Section 4.2 is dedicated to model formulation. In Section 4.3, the ZAGOLLMax semiparametric regression models are presented as well as some inferential results. A simulation study is presented for the ZAGOLLMax semiparametric regression model in Section 4.5. In Section 4.4 we discuss the diagnostic and residuals for the ZAGOLLMax semiparametric regression model. Results of an application to a real data set are reported in Section 4.6. In Section 4.7, we end up with some general remarks.

4.2 Model formulation

A random variable Y has the GOLLMax distribution if its cumulative distribution function (cdf) and probability density function (pdf) are (for $y > 0$)

$$H(y;\nu,\sigma,\mu) = \frac{\gamma_1^{\nu\,\sigma}(3/2, y^2/\mu^2)}{\gamma_1^{\nu\sigma}(3/2, y^2/\mu^2) + [1-\gamma_1^{\sigma}(3/2, y^2/\mu^2)]^{\nu}}. \tag{4.1}$$

and

$$h(y;\nu,\sigma,\mu) = \frac{4\,\nu\,\sigma\,y^2}{\sqrt{\pi}\,\mu^3}\,\exp\left(-\frac{y^2}{\mu^2}\right)\frac{\gamma_1^{\nu\sigma-1}(3/2, y^2/\mu^2)\left[1-\gamma_1^{\sigma}(3/2, y^2/\mu^2)\right]^{\nu-1}}{\left\{\gamma_1^{\nu\sigma}(3/2, y^2/\mu^2) + [1-\gamma_1^{\sigma}(3/2, y^2/\mu^2)]^{\nu}\right\}^2}, \tag{4.2}$$

respectively, where $\nu > 0$ and $\sigma > 0$ are two extra shape parameters and $\mu > 0$ is scale parameter.

Henceforth, if Y is a random variable with cdf (4.1), we write $Y \sim \text{GOLLMax}(\nu,\sigma,\mu)$.

The GOLLMax distribution can be simulated by inverting (4.1). The quantile function (qf) of Y, is given

$$y = Q_{\text{Max}}\left(\left[\frac{\left(\frac{u}{1-u}\right)^{\frac{1}{\nu}}}{1+\left(\frac{u}{1-u}\right)^{\frac{1}{\nu}}}\right]^{\frac{1}{\sigma}}\right), \tag{4.3}$$

where $Q_{\text{Max}}(u) = G^{-1}(\mu; u)$ is the qf of the Maxwell distribution.

There are situations where continuous data can include a high percentage of zeros. In these situations, continuous distributions can not be used, in this research we will assume that the continuous component is described by the GOLLMax distribution, since this model is very flexible to describe the behavior of the phenomenon under study. Meanwhile, the discrete component (mass point) will be described by means of a degenerate distribution at zero point. According to Heller *et al.* (2006), Ospina and Ferrari (2012) and Hashimoto *et al.* (2018), data that contain excessive zeros can be analyzed by a mixture of two distributions: a continuous distribution defined by the pdf $h(y)$ (with positive support) and cdf $H(y)$ and a degenerate distribution at zero, i.e. a model whose mixed discrete-continuous probability and distribution functions. Thus we introduce the zero adjusted generalized odd log-logistic Maxwell (ZAGOLLMax) model defined by mixed discrete-continuous probability and distribution functions are

$$f(y;\mu,\sigma,\nu,\tau) = \begin{cases} \tau & \text{if} \quad y = 0, \\ \frac{(1-\tau)\,4\,\nu\,\sigma\,y^2}{\sqrt{\pi}\,\mu^3}\,\exp\left(-\frac{y^2}{\mu^2}\right)\frac{\gamma_1^{\nu\sigma-1}(3/2,y^2/\mu^2)\left[1-\gamma_1^{\sigma}(3/2,y^2/\mu^2)\right]^{\nu-1}}{\left\{\gamma_1^{\nu\sigma}(3/2,y^2/\mu^2)+\left[1-\gamma_1^{\sigma}(3/2,y^2/\mu^2)\right]^{\nu}\right\}^2} & \text{if} \quad y > 0, \end{cases} \tag{4.4}$$

and

$$F(y) = I_{\{y=0\}}(y)\,\tau + I_{\{y>0\}}\frac{(1-\tau)\,\gamma_1^{\nu\,\sigma}(3/2,y^2/\mu^2)}{\gamma_1^{\nu\sigma}(3/2,y^2/\mu^2)+\left[1-\gamma_1^{\sigma}(3/2,y^2/\mu^2)\right]^{\nu}} \tag{4.5}$$

respectively and we use the indicator function $I_A(y)$ is one if $y \in A$ and zero if $y \in A$.

Some plots of the ZAGOLLMax pdf for selected parameter values are displayed in Figure 4.3. A characteristic of the distribution is that its pdf can be unimodal, bimodal, among others, depending basically on the parameter values.

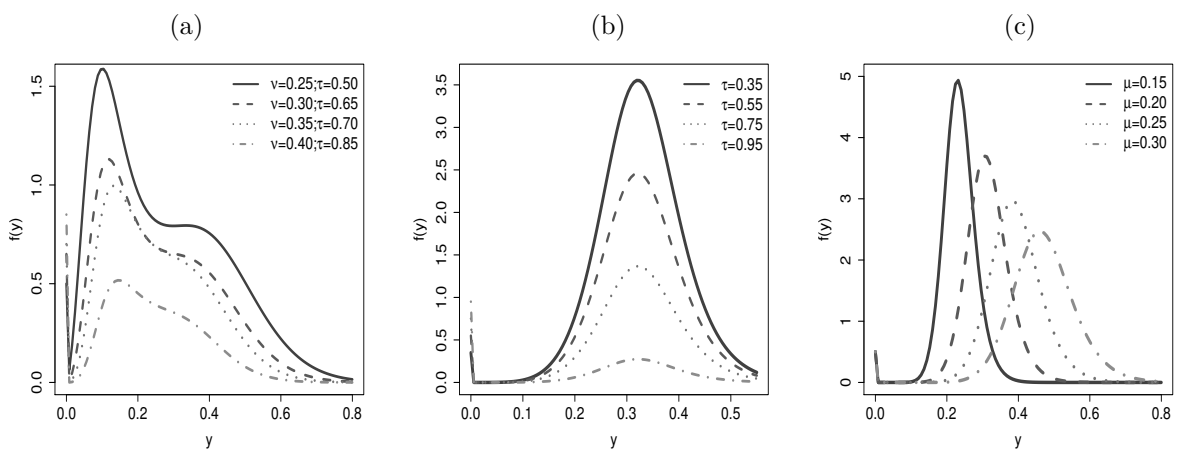

Figure 4.3. Plots of the ZAGOLLMax density for some parameter values. (a) For different values of the ν and τ with $\mu = 0.15$ and $\sigma = 3.45$ (b) For different values of the τ with $\mu = 0.30$, $\sigma = 1.00$ and $\nu = 2.00$. (c) For different values of the μ with $\sigma = 3.45$, $\nu = 1.45$ and $\tau = 0.50$.

4.3 The ZAGOLLMax semiparametric regression model

In many practical applications, the response variable Y can be influenced by explanatory variables. In this work we consider that these covariables may have linear and nonlinear effects in relation to the response variable. Therefore, to study the effect of these explanatory variables on the response

variable and the shape parameters, we can consider a regression model, where the response variable has the GOLLMax distribution given by (4.5). Let $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)^T$ denote the parameter vector of this density. Further, let $\boldsymbol{\theta}_i = (\mu_i, \sigma_i, \nu_i, \tau_i)^T$ be a parameter vector related to the ith response variable. We consider that the independent observations y_i conditional on $\boldsymbol{\theta}_i$ (for $i = 1, \ldots, n$) has pdf $f(y_i; \boldsymbol{\theta}_i)$.

Therefore we will define the ZAGOLLMax semiparametric regression model regression models. We define the semiparametric systematic component

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\sigma} \\ \boldsymbol{\nu} \\ \boldsymbol{\tau} \end{pmatrix} = \begin{pmatrix} g_1\Big(\mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1})\Big) \\ g_2\Big(\mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=2}^{J_2} h_{j2}(\mathbf{x}_{j2})\Big) \\ g_3\Big(\mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=3}^{J_3} h_{j3}(\mathbf{x}_{j3})\Big) \\ g_3\Big(\mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=4}^{J_4} h_{j4}(\mathbf{x}_{j4})\Big) \end{pmatrix}, \tag{4.6}$$

where $h_{jk}(\mathbf{x}_{jk})$ are smooth functions of the covariables $\mathbf{x}_{jk}$ for $k = 1, \ldots, 4$ and $j = 1, \ldots, J_k$. In this chapter, the approximation of $h_{jk}(\cdot)$ is by cubic spline. In the `gamlss` package of the `R` software, such smoothing functions are expressed as random effects, i.e. $h_{jk}(\cdot) = Z_j\boldsymbol{\gamma}_j$, where Z_j is the $(n \times q_j)$ known basis design matrix and $\boldsymbol{\gamma}_j$ is the q_j-dimensional unknown vector of parameters.

We shall consider the logarithmic link function for $g_k(\cdot)$ ($k = 1$, 2 and 3) and the logit link function for $g_4(\cdot)$. However, in this work we assume a systematic regression structure, only in the μ_i and τ_i parameters, considering for σ_i, $\log(\sigma_i) = \beta_{02}$ and ν_i, $\log(\nu_i) = \beta_{03}$ constants, respectively. Thus, the follow systematic components are given by

$$\log(\mu_i) = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1}), \qquad \text{and} \qquad \text{logit}(\tau_i) = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} h_{j4}(\mathbf{x}_{j4}), \tag{4.7}$$

reversing these functions we directly recover μ_i and τ_i as follows

$$\mu_i = \exp\Big(\mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1})\Big), \qquad \text{and} \qquad \tau_i = \frac{\exp\Big(\mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=4}^{J_4} h_{j4}(\mathbf{x}_{j4})\Big)}{1 + \exp\Big(\mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=4}^{J_4} h_{j4}(\mathbf{x}_{j4})\Big)},$$

for $i = 1, \ldots, n$.

For the semiparametric model with structure by (4.7), the fixed and random effects $\boldsymbol{\theta} = (\nu, \sigma, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ and $\boldsymbol{\eta}$, respectively, are estimated by maximizing the penalized log-likelihood function

$$\begin{aligned} l(\boldsymbol{\omega}) \;=\; & r\log\left(\frac{4\,\nu\,\sigma}{\sqrt{\pi}}\right) + \sum_{i:y_i=0}\log(\tau_i) + \sum_{i:y_i>0}\log(1-\tau_i) + \sum_{i:y_i>0}\log\left(\frac{y_i^2}{\mu_i^3}\right) - \sum_{i:y_i>0}\left(\frac{y_i}{\mu_i}\right)^2 + \\ & (\nu\,\sigma - 1)\sum_{i:y_i>0}\log\left[\gamma_1(3/2, y_i^2/\mu_i^2)\right] + (\nu - 1)\sum_{i:y_i>0}\log\left[1 - \gamma_1^{\sigma}(3/2, y_i^2/\mu_i^2)\right] - \\ & 2\sum_{i:y_i>0}\log\left\{\gamma_1^{\nu\,\sigma}(3/2, y_i^2/\mu_i^2) + [1 - \gamma_1^{\sigma}(3/2, y_i^2/\mu_i^2)]^{\nu}\right\} - \frac{1}{2}\sum_{j=1}^{J}\lambda_j\boldsymbol{\eta}_j^T\mathbf{P}_j\boldsymbol{\eta}_j^T, \end{aligned} \tag{4.8}$$

λ_j is the unknown smoothing parameter and P_j is a symmetric matrix that may depend on a vector of smoothing parameters and chosen number of knots. The solution of (4.8) corresponds to the cubic smoothing spline with equidistant knots for distinct x-variable values (see more details in Green and Silverman, 1993, Ruppert, *et al.* 2003 and Wood, 2017). Another measure of interest are the effective degrees of freedom, df^*, relative to the non-parametric component. The smoothing parameters can be fixed or estimated from the data. Some methods are proposed in the literature, for example, by the generalized cross-validation method (see Wood, 2017). However, in this work, we consider in simulations and applications the default $df^* = 3$.

The proposed model ZAGOLLMax was implemented numerically in the structure `gamlss` package in `R` software. The functions used are presented in Appendix B. We use the maximization by `RS` algorithm described by Stasinopoulos and Rigby (2007) and Stasinopoulos *et al.* (2017). The `cs()` function is used to assign the arguments to make the adjustment via `gamlss`. Thus, in Section 4.6, the `cs(·)` function is denoted by $cs(\mathbf{x}_{jk}, df^* = 3)$. The effective degree of freedom for structure of the regression in μ considering an explanatory variable x, for example, is given by $df_\mu = df^* + 2$ where other two additional degrees of freedom are in relation to the linear terms (see Voudouris *et al.* (2012)). Finally, we have that the total freedom degree of the adjusted model, represented by df, collectively considers the additive terms and the parametric terms, i.e, $df = df_\mu + df_\sigma + df_\nu + df_\tau$, which are the freedom degrees used to model μ, σ, ν and τ, respectively.

4.4 Checking model

The selection of the appropriate distribution is performed in two steps, the adjustment phase considering the marginal analysis (only response variable) and modeling with regression structure (considering the complete model), according to the chosen model, we proceed to the diagnostic step. In the first step, we use the global deviance (GD), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The GD is given by $GD = -2\,l(\hat{\boldsymbol{\theta}})$, where $l_p(\hat{\boldsymbol{\theta}})$ is the total log-likelihood function and the AIC and BIC criterion are obtained by $AIC = GD + 2\,df$ and $AIC = GD + \log(n)\,df$, where df is the total effective degrees of freedom of the fitted model. The model with the smallest values of these criteria is then selected.

In the diagnostic step, the model assumptions and the presence of outlying observations are checked. We can use the diagnostic tools in the `gamlss` package. Some of the diagnostic techniques used in this work are described below. We will consider the normalized randomized quantile residuals (Dunn and Smyth, 1996). These residuals are determined by $\hat{r}_i = \Phi^{-1}(\hat{u}_i)$, where $\Phi^{-1}(\cdot)$ is the inverse cdf of the standard normal distribution. We have that $\hat{u}_i$ is a random variable with uniform distribution in interval $(a_i, b_i]$, where $a_i = \hat{F}(y_i - 1|\hat{\boldsymbol{\theta}})$ and $b_i = \hat{F}(y_i|\hat{\boldsymbol{\theta}})$ are the fitted distribution function (4.5). Considering that the model was specified correctly, these residues will have standard normal distribution. That said, several usual diagnostic techniques can be performed to verify the model.

We also use Worm Plots (WP) as the technique to check the adjustment quality. Worm plots of the residuals were introduced by Buuren and Fredriks (2001). The general idea of these plots is to identify regions (intervals) of an explanatory variable within which the model does not fit adequately the data. The WP is a detrended normal QQ-plot of the residuals. Model inadequacy is indicated when many points plotted lie outside the pointwise 95% confidence bands or when the points follow a systematic shape. For example, the interpretations of the shapes of the WP are: a vertical shift, a slope, a parabola or an S shape, thus indicating a misfit in the mean, variance, skewness and excess kurtosis of the residuals, respectively.

For the verification of the ability of the model to reproduce the current data, we construct simulated envelopes considering the complete model (Atkinson, 1985), only the continuous component and only the discrete part of the regression model. In this way, these envelopes help in a better interpretation of the normal probability graph with based in the residues.

Similar to the complete model the envelopes for the continuous component are constructed simulating values of the model (4.1), the residues are obtained by fitting the model (4.7). For the discrete component we consider a binary variable in which it receives the value 1 if $y_i = 0$ and 0 if $y_i > 0$, the residues are obtained considering the steps above with the binomial model.

4.5 Simulation study for ZAGOLLMax semiparametric regression model

In this section, we examine the performance of the ZAGOLLMax semiparametric regression model by means of a Monte Carlo simulation study under two scenarios. Various simulations are conducted for different sample sizes ($n = 200, 400, 900$) using the `R` software with `gamlss` packages by `RS` method.

In this study we present and compare the results, adjusting the proposed GOLLMax model (parametric and semiparametric). The following three scenarios with regression structure (4.7) are presented:

- In the first scenario, we consider the following structure for regression model with parameter $\mu_i = \exp\{h_{11}(x_{i1}) + \beta_{21}x_{i2} + \beta_{31}x_{i3}\}$, where $h_{11}(x_{i1}) = 0.45[\sin(\pi\, x_{i1})]$ with the functional shape presented in Figure 4.4a, $\sigma = \exp\{\beta_{02}\}$, $\nu = \exp\{\beta_{03}\}$ and $\tau_i = \text{logit}\{\beta_{04} + \beta_{14}x_{i1}\}$. The values associated to the coefficients are: $\beta_{21} = 0.20$, $\beta_{31} = -0.35$, $\beta_{03} = 0.45$, $\beta_{04} = 1.20$, $\beta_{04} = -0.95$ and $\beta_{14} = 0.30$, with a percentage of zeros approximately 0.35.

- In the second scenario, we consider the following structure for regression model with parameter $\mu_i = \exp\{h_{11}(x_{i1}) + \beta_{21}x_{i2} + \beta_{31}x_{i3}\}$, where $h_{11}(x_{i1}) = \sin(x_{i1})$ with the functional shape presented in Figure 4.4b, $\sigma = \exp\{\beta_{02}\}$, $\nu = \exp\{\beta_{03}\}$ and $\tau_i = \text{logit}\{\beta_{04} + \beta_{14}x_{i1}\}$. The values associated to the coefficients are: $\beta_{21} = 0.20$, $\beta_{31} = -0.35$, $\beta_{02} = 0.45$, $\beta_{03} = 1.20$, $\beta_{04} = -0.65$ and $\beta_{14} = 1.30$, with a percentage of zeros approximately 0.70.

- In the third scenario, we consider the idea presented in Ramires *et al.* (2018) for nonlinear effects in regression models with long-term survival. Thus, the following structure for regression model with parameter $\mu_i = \exp\{h_{11}(x_{i1}) + \beta_{21}x_{i2} + \beta_{31}x_{i3}\}$, where $h_{11}(x_{i1}) = \sin(x_{i1})$ with the functional shape presented in Figure 4.4b, $\sigma = \exp\{\beta_{02}\}$, $\nu = \exp\{\beta_{03}\}$ and structure of the regression model for parameter $\tau_i = \text{logit}\{h_{14}(x_{i4})\}$ with the functional shape presented in Figure 4.4c. For each level of x_{i4}, it was generated a sample size of length n_i, so that $n = \sum_{i=1}^{10} n_i$. The fixed values of τ, for each value of the x_4, are given in Table 4.5, associated to the coefficients are: $\beta_{21} = 0.20$, $\beta_{31} = -0.35$, $\beta_{02} = 0.45$, $\beta_{03} = 1.20$, $\beta_{04} = -0.65$ and $\beta_{14} = 1.30$, with a percentage of zeros approximately 0.70.

Table 4.1. Fixed values of the τ parameter of each level of the x_{i4} explanatory variable.

τ	0.20	0.25	0.35	0.40	0.45	0.45	0.40	0.35	0.25	0.20
x_4	1	2	3	4	5	6	7	8	9	10

We assume that the explanatory variables for the two scenarios are $x_{i1} \sim \text{Uniform}(0, 2.5)$, $x_{i2} \sim \text{Normal}(0, 5.50)$ and $x_{i3} \sim \text{Uniform}(0, 1)$, respectively, for $i = 1, \ldots, n$.

To generate the random values of the proposed model with zero proportion, we present a brief script:

i. Generate the vector of proportions by $p \sim U(0, 1)$ of size n.

ii. Create a function $W = F(q, \mu = \mu[i], \sigma = \sigma[i], \nu = \nu[i], \tau = \tau[i]) - p[i]$, where q is the quantil evaluated in the distribution function (4.3) and p is the probability to be assessed by the qf.

iii. Make the condition to choose the zeros and the continuous values
- if $(\tau[i] >= p[i])$ $q[i] = 0$

- else uniroot(W, c(lower[i], upper[i]))$root
- if ($q[i] >= upper[i]$) warning("q is at the upper limit, increase the upper.limit").

iv. Thus, $y = q$ where $y \sim$ ZAGOLLMax$(\mu_i, \sigma, \nu, \tau_i)$.

The samples can be generated in R using the code rZAGOLLMax$(n, \mu, \sigma, \nu, \tau)$.

(a) (b) (c)

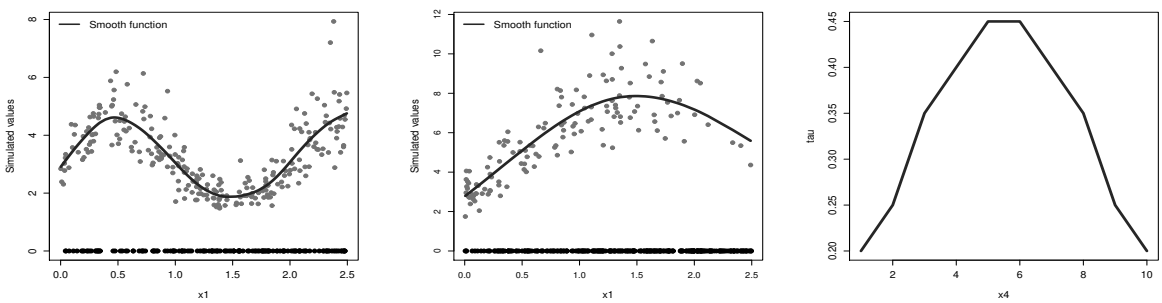

Figure 4.4. Plots of the simulation values. (a) Nonlinear effect for scenario 1. (b) Nonlinear effect for scenarios 2 and 3. (c) Nonlinear effect for scenario 3.

For each of the $1,000$ simulations, the average estimates (AEs), biases and mean square errors (MSEs) are calculated. The results are reported in Tables 4.2, 4.3 and 4.4 for the parametric and semiparametric models. Based on the simulation results in Tables 4.2, 4.3 and 4.4, we are interested in verifying how much the inclusion of an additive term affects in the estimations of the other fixed parameters. For semiparametric model, we verify that the MSEs of the maximum likelihood estimates (MLEs) of β_{21}, β_{31}, β_{02}, β_{03}, β_{04} and β_{14} for scenario 1, 2 and 3 decay towards zero when the sample size n increases, as usually expected under first-order asymptotic theory. The mean estimates of the parameters tend to be closer to the true parameter values when n increases. However, for the parametric model, such measures do not exhibit the same behavior.

In relation to the behavior of the nonlinear effects in the simulations (scenarios 1, 2 and 3), the Figures 4.5, 4.6 and 4.7 display the generated and fitted effects for the parametric and semiparametric models. We also present in this figure the box-plots of the GD, AIC and BIC statistics obtained in 1,000 simulations for both models. We can note that the nonlinear effects are very close to the true shape as shown in the Figure 4.4, when the sample size increases. Further, we can conclude that the semiparametric model presents the lowest values of GD, AIC and BIC statistics, indicating that it is the most suitable model for simulated data in the presence of non-linear effects.

4.6 Data analysis

The set of microbiological data was obtained from an experiment conducted in field conditions in the municipality of Novo Horizonte, state of São Paulo, Brazil (21º29'42"S - 49º11'23"W; altitude of 462 meters).

Agricultural experiments conducted under field conditions generally adopt a randomized block design with the purpose of reducing the effects of environmental heterogeneity. In this respect, the experiment was conducted in a randomized block design with four repetitions, where the sugarcane cultivar CTC 24 was used. Each treatment consisted of seven rows of plants, spaced 1.5 meters apart, with

Table 4.2. The AEs, biases and MSEs for the parametric and semiparametric ZAGOLLMax regression models based on 1,000 simulations for scenario 1.

	Semiparametric				Parametric			
n	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.199	-0.001	0.001	β_{21}	0.199	-0.001	0.002
	β_{31}	-0.364	-0.014	0.002	β_{31}	-0.439	-0.089	0.012
200	β_{02}	0.417	-0.033	0.001	β_{02}	0.898	0.448	0.281
	β_{03}	1.118	-0.082	0.010	β_{03}	-0.088	-1.282	1.674
	β_{04}	-0.957	-0.007	0.076	β_{04}	-0.957	-0.007	0.076
	β_{14}	0.300	0.000	0.040	β_{14}	0.300	0.000	0.040
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.206	0.006	0.000	β_{21}	0.202	0.002	0.001
	β_{31}	-0.366	0.016	0.001	β_{31}	-0.354	-0.004	0.002
400	β_{02}	0.416	-0.034	0.001	β_{02}	1.440	0.990	1.036
	β_{03}	1.089	-0.111	0.014	β_{03}	-0.430	-1.673	2.673
	β_{04}	-0.939	0.011	0.048	β_{04}	-0.939	0.048	0.048
	β_{14}	0.293	-0.007	0.021	β_{14}	0.293	0.021	0.021
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.202	0.002	0.000	β_{21}	0.187	-0.013	0.001
	β_{31}	-0.346	0.004	0.000	β_{31}	-0.339	0.011	0.001
900	β_{02}	0.417	-0.033	0.001	β_{02}	1.005	0.555	0.327
	β_{03}	1.095	0.105	0.012	β_{03}	-0.176	-1.376	0.900
	β_{04}	-0.948	0.002	0.021	β_{04}	-0.948	0.002	0.021
	β_{14}	0.294	-0.006	0.010	β_{14}	0.294	-0.006	0.010

Table 4.3. The AEs, biases and MSEs for the parametric and semiparametric ZAGOLLMax regression models based on 1,000 simulations for scenario 2.

	Semiparametric				Parametric			
n	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.204	0.004	0.001	β_{21}	0.211	0.011	0.002
	β_{31}	-0.351	-0.001	0.003	β_{31}	-0.309	0.041	0.007
200	β_{20}	0.433	-0.017	0.001	β_{02}	0.335	-0.115	0.018
	β_{03}	1.293	0.093	0.020	β_{03}	0.848	-0.352	0.139
	β_{04}	-0.660	-0.010	0.113	β_{04}	-0.660	-0.010	0.113
	β_{14}	1.321	0.021	0.083	β_{14}	1.321	0.021	0.083
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.199	-0.001	0.000	β_{21}	0.217	0.017	0.001
	β_{31}	-0.349	0.001	0.001	β_{31}	-0.364	-0.014	0.003
400	β_{02}	0.426	-0.024	0.001	β_{02}	0.315	-0.135	0.022
	β_{03}	1.245	0.045	0.008	β_{03}	0.781	-0.419	0.183
	β_{04}	-0.664	-0.014	0.047	β_{04}	-0.664	-0.014	0.047
	β_{14}	1.316	0.016	0.033	β_{14}	1.316	0.016	0.033
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.200	0.000	0.000	β_{21}	0.207	0.007	0.000
	β_{31}	-0.350	0.000	0.001	β_{31}	-0.336	0.014	0.001
900	β_{02}	0.424	-0.026	0.001	β_{02}	0.301	-0.149	0.027
	β_{03}	1.224	0.024	0.003	β_{03}	0.768	-0.432	0.192
	β_{04}	-0.662	-0.012	0.020	β_{04}	-0.662	-0.012	0.020
	β_{14}	1.314	0.014	0.015	β_{14}	1.314	0.014	0.015

a length of 20 m, resulting in an area of $30m^2$. Therefore, the area of each treatment was approximately $210m^2$. In this study, we consider data from four blocks, each containing seven treatments.

Therefore, three types of compost were assessed (non-enriched compost, compost enriched with

Table 4.4. The AEs, biases and MSEs for the parametric and semiparametric ZAGOLLMax regression models based on 1,000 simulations for scenario 3.

	Semiparametric				Parametric			
n	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.202	0.002	0.001	β_{21}	0.222	0.022	0.00'
	β_{31}	-0.350	0.000	0.001	β_{31}	-0.350	0.000	0.002
200	β_{20}	0.426	-0.024	0.001	β_{02}	0.330	-0.120	0.069
	β_{03}	1.244	0.044	0.007	β_{03}	0.589	-0.611	0.388
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.198	-0.002	0.000	β_{21}	0.187	0.013	0.001
	β_{31}	-0.353	0.003	0.001	β_{31}	-0.423	-0.073	0.007
400	β_{02}	0.425	-0.025	0.001	β_{02}	1.446	0.996	1.546
	β_{03}	1.227	0.027	0.003	β_{03}	0.004	-1.196	1.551
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.201	0.001	0.000	β_{21}	0.195	-0.005	0.000
	β_{31}	-0.348	0.002	0.000	β_{31}	-0.350	0.000	0.001
900	β_{02}	0.424	-0.026	0.001	β_{02}	1.288	0.838	0.832
	β_{03}	1.219	0.019	0.001	β_{03}	0.077	-1.123	1.291

(a) (b) (c)

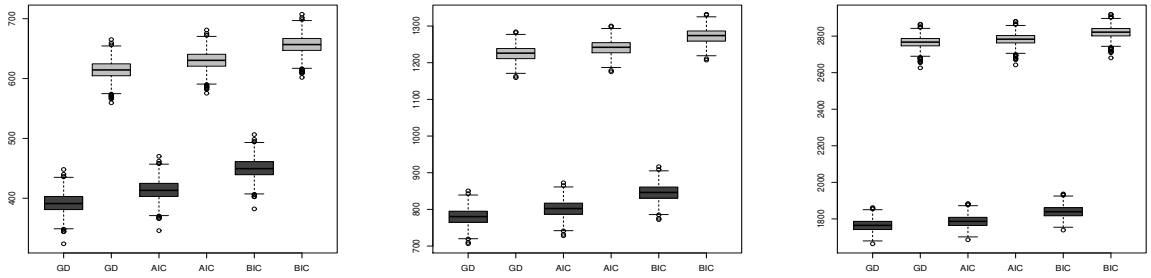

Figure 4.5. The fitted ZAGOLLMax semiparametric and parametric regression model and box plots for GD, AIC and BIC measures for scenario 1, with the black line is the mean of the fitted values. (a) n=200. (b) n=400. (c) n=900.

apatite and compost enriched with phosphorite). These composts were or were not inoculated in the field with phosphate solubilizing bacteria, for a total of 6 treatments. Finally, one additional treatment was established, where the plants only received the mineral fertilizer commonly used in conventional large-scale sugarcane crops, including phosphorus in the form of triple superphosphate, summing up seven treatments in all (Table 4.5).

(a) (b) (c)

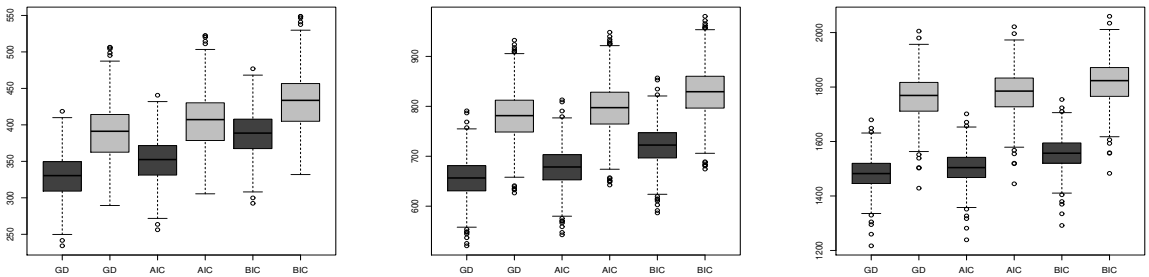

Figure 4.6. The fitted ZAGOLLMax semiparametric and parametric regression model and box plots for GD, AIC and BIC measures for scenario 2, with the black line is the mean of the fitted values. (a) n=200. (b) n=400. (c) n=900.

The soil samples were collected at six and 12 months after planting, corresponding to the first cycle of sugarcane growth. Therefore, we consider the factor period in the analysis of the data.

Table 4.5 presents the description of the study treatments:

Table 4.5. Description of treatments.

Treatments	Description	Inoculation with PSB [1]	ACA (tons.ha^{-1})
Treat 1	Mineral (Control)	-	-
Treat 2	Compost	Without	19,00
Treat 3	Compost	With	19,00
Treat 4	Compost with apatite	With	9,75
Treat 5	Compost with apatite	Without	9,75
Treat 6	Compost with phosphorite	Without	9,75
Treat 7	Compost with phosphorite	With	9,75

[1]Phosphate solubilizing bacteria.
[2]Amount of compost applied in (tons.ha^{-1}).

Here we provide a brief description of the response variable (y=`T-RFLP`). After collection, the soil samples were stored at -80°C for subsequent extraction of total DNA. To ascertain the structure of the fungal community, the terminal restriction fragment length polymorphism technique was used (y). In this technique, aliquots of the extracted DNA are submitted to amplification of specific fragments by polymerase chain reaction (PCR). The primers used were ITS1 (5'-CTTGGTCATTTAGAGGAAGTAA-3') marked with 5-carboxyfluorescein(5-FAM) (Gardes and Bruns, 1993) and ITS4 (5'-TCCTCCGCTTATTG ATATC-3') (White *et al.* 1990). At the end of the amplification, the samples were sequenced, generating

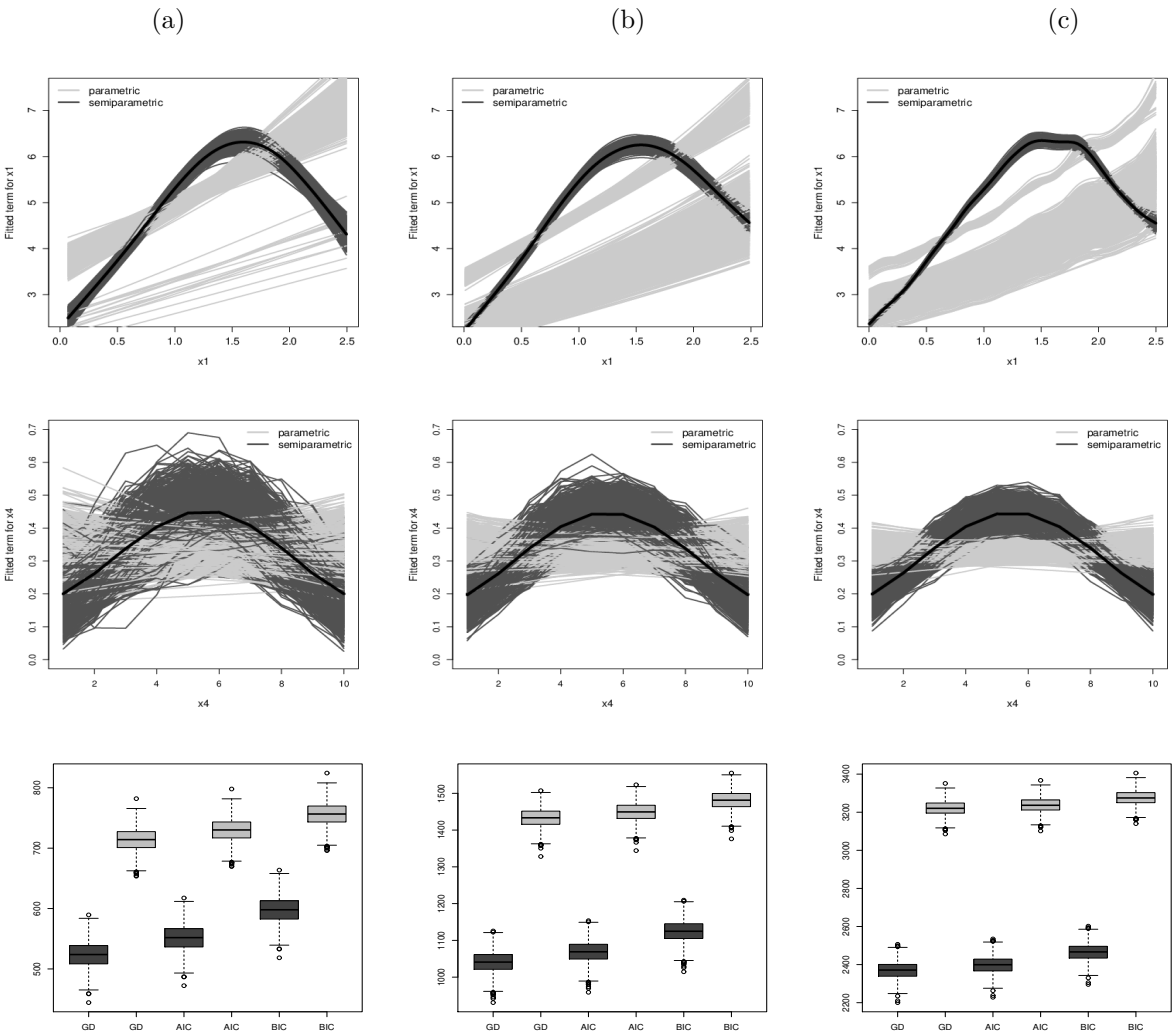

Figure 4.7. The fitted ZAGOLLMax semiparametric and parametric regression model and box plots for *GD*, *AIC* and *BIC* measures for scenario 3, with the black line is the mean of the fitted values. (a) n=200. (b) n=400. (c) n=900.

an electropherogram, from which the peak areas were used to compose the data matrix employed in this study. A cutoff line of 50 fluorescence units was applied to avoid background noise (such as noise generated by the equipment). With these results, a profile was obtained with varied peaks, also called restriction fragments `T-RFs` (Liesack and Dunfield, 2004). The profile `T-RFs` differentiates between the samples, which may or may not be present, so that there is greater or lesser prevalence in the proportion of zeros.

Besides the structure of the fungal community, the following variables were quantified:

- Abundance of bacterial genes (x_1 = `16S RNAr`). The genes `16S RNAr` were quanitied by using the primers Eub338 (5' - CCTACG GGA GGC AGC AG-3') (Muyzer *et al.* 1993) and Eub518 (5' - ATTACC GCG GCT GCT GG - 3') (Muyzer *et al.* 1993), generating a fragment of 193 base pairs for bacteria.

- The total fungi (x_2 = `ITS`) were obtained from specific fragments for each `ITS` region. To obtain

the data, the primers used were ITS1f (5'-CTTGGTCATTTAGAGGAAGTA A-3') (Gardes and Bruns, 1993) and 5.8 s (5'-CGCTGCGTTCTTCATCG-3') (Gardes and Bruns, 1993), generating a fragment with 300 base pairs. The gene quantification was performed with the StepOne^{TM} Real-Time PCR System (Applied Biosystems, Life Technologies), with the SYBR^{R} GreenER^{TM}.

- Abundance of the gene ($x_3 = phoD$), related to the synthesis of phosphatase, an enzyme related to the availability of phosphorus in the soil, was quantified using the primers ALPS-F730 (5' - CAG TGG GAC GAC CAC GAG GT-3') (Sakurai *et al.* 2008) and ALPS-R1101 (5' -GAG GCC GAT CGG CAT GTC G - 3') (Sakurai *et al.* 2008), generating a fragment of 370 base pairs.

4.6.1 Descriptive and marginal analyses

First, we present an exploratory and marginal analysis of the microbiological data for the `T-RFLP` variable. Table 4.6 provides the mean, median, variance, standard deviation, asymmetry, kurtosis, minimum, maximum and proportion of zeros measures. In this table, the highest values in relation to the mean, median and proportion of zeros are highlighted by treatments 2 and 3 in relation to the others and especially in relation to treatment 1 control.

Table 4.6. Descriptive statistics for the y variable by blocks, treatments and periods.

Variables	Mean	Median	Variance	SD	Skewness	Kurtosis	Min.	Max.	% zero
Block 1	0.909	0.917	0.097	0.312	0.084	-0.530	0.197	1.680	0.200
Block 2	0.808	0.744	0.158	0.397	0.235	-1.246	0.117	1.666	0.202
Block 3	0.686	0.629	0.127	0.356	0.695	-0.174	0.154	1.844	0.182
Block 4	0.629	0.533	0.134	0.367	0.640	-0.255	0.018	2.004	0.173
Treat 1	0.685	0.627	0.102	0.320	0.572	-0.313	0.018	1.572	0.100
Treat 2	0.862	0.880	0.133	0.351	0.437	0.062	0.263	2.004	0.118
Treat 3	0.812	0.835	0.141	0.376	0.391	-0.509	0.193	1.844	0.115
Treat 4	0.759	0.765	0.176	0.420	0.160	-1.205	0.067	1.731	0.113
Treat 5	0.646	0.577	0.141	0.375	0.626	-0.548	0.084	1.666	0.101
Treat 6	0.765	0.745	0.174	0.417	0.128	-1.363	0.129	1.531	0.116
Treat 7	0.693	0.638	0.123	0.350	0.499	-0.835	0.172	1.476	0.103
Period 1	0.795	0.776	0.127	0.357	0.287	-0.565	0.018	2.004	0.394
Period 2	0.679	0.606	0.146	0.382	0.563	-0.671	0.067	1.844	0.372
y	0.732	0.664	0.141	0.375	0.408	-0.695	0.018	2.004	0.767

Table 4.7 presents some descriptive data means, medians, variances, standard deviations, asymmetries, kurtoses, minima, maxima of the variables x_1, x_2 and x_3.

Table 4.7. Descriptive statistics for the x_1, x_2 and x_3 variables.

Variables	Mean	Median	Variance	SD	Skewness	Kurtosis	Min.	Max.
x_1	9.748	9.755	0.011	0.105	0.436	0.759	9.560	10.070
x_2	6.023	5.995	0.061	0.247	-0.024	0.216	5.300	6.630
x_3	6.663	6.640	0.023	0.154	1.325	2.538	6.420	7.250

The first step before fitting a regression model is to select a suitable distribution for the response variable. In addition to the distributions GOLLMax, we consider two others: gamma (GA) and inverse gaussian (IG) distributions.

The GA distribution, denoted by GA(μ, σ), is defined by (for $y > 0$)

$$f(y) = \frac{y^{(\frac{1}{\sigma^2}-1)}}{\Gamma\left(1/\sigma^2\right)\left(\mu\sigma^2\right)^{\frac{1}{\sigma^2}}} \exp\left\{-\frac{y}{\mu\sigma^2}\right\}, \tag{4.9}$$

were $\mu > 0$ and $\sigma > 0$, with measures $E(y) = \mu$ and $Var(y) = \mu^2\sigma^2$ are the mean and variance of the gamma distribution, respectively.

The IG distribution, denoted by IG(μ, σ), is defined by

$$f(y) = \frac{1}{\sqrt{2\pi y^3\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\mu^2\sigma^2\, y}\right\}, \tag{4.10}$$

where $y > 0$, $\mu > 0$ and $\sigma > 0$, with measures $E(y) = \mu$ and $Var(y) = \mu^3\sigma^2$ are the mean and variance of the inverse gaussian distribution, respectively.

The zero inflated or zero adjusted gamma (ZAGA) and inverse gaussian (ZAIG) distributions are a mixture of such zero degenerate distributions similarly to that presented in equation (4.5). Thus, we denote by ZAGA(μ, σ, τ) and ZAIG(μ, σ, τ) with the parameter τ that represents the ratio of zero occurrence.

The fitted models to the `T-RFLP` variable are compared by means of the GD, AIC and BIC using the `gamlss` package of the `R` software. In this analysis we consider the `T-RFLP` variable without zeros and with zeros, respectively. In Tables 4.8 and 4.9, we give the values of GD, AIC and BIC. The smaller these values more appropriate the model. For this study, the GOLLMax distribution yields the lowest values.

Table 4.8. MLEs of the parameters of the GOLLMax, GA and IG models for `T-RFLP` variable, the corresponding SEs (given in parentheses) and statistics: GD, AIC and BIC.

Model	$\log(\alpha)$	$\log(\sigma)$	$\log(\nu)$	GD	AIC	BIC
GOLLMax	-0.755	0.880	-0.844	571.8	577.8	591.8
	(0.060)	(0.194)	(0.123)			
GA	-0.311	-0.598		624.6	628.6	637.9
	(0.019)	(0.024)				
IG	-0.311	-0.220		770.4	774.4	783.7
	(0.024)	(0.025)				

Table 4.9. MLEs of the parameters of the ZAGOLLMax, ZAGA and ZAIG models for `T-RFLP` variable, the corresponding SEs (given in parentheses) and statistics: GD, AIC and BIC.

Model	$\log(\mu)$	$\log(\sigma)$	$\log(\nu)$	logit(τ)	GD	AIC	BIC
ZAGOLLMax	-0.755	0.880	-0.844	1.194	4215.5	4223.5	4248.0
	(0.060)	(0.194)	(0.123)	(0.040)			
ZAGA	-0.311	-0.598		1.194	4268.2	4274.2	4292.6
	(0.019)	(0.024)		(0.040)			
ZAIG	-0.311	-0.220		1.194	4414.0	4420.0	4438.4
	(0.024)	(0.025)		(0.040)			

Figure 4.8 reveals the presence of asymmetry and bimodality in the response variable `T-RFLP` without zeros. In order to assess if the model is appropriate, plots of the fitted GOLLMax, GA and IG density functions and the histogram are displayed in Figure 4.8a. Plots of the fitted GOLLMax, GA and IG cumulative and empirical cdfs are given in Figure 4.8b. They also reveal that the GOLLMax distribution provides a good fit to these data.

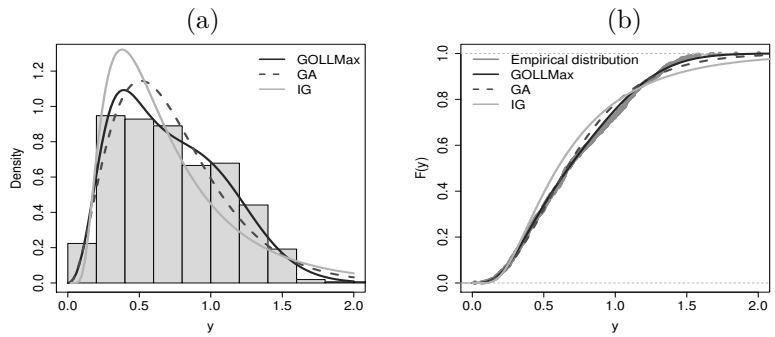

Figure 4.8. Plots for `T-RFLP` variable without the zeros. (a) Histogram with estimated density functions of the GOLLMax, GA and IG models. (b) Empirical distribution with estimated cumulative functions of the GOLLMax, GA and IG models.

In the presence of zeros in the response variable `T-RFLP`, Figure 4.9 reveals the presence of asymmetry and bimodality in the response variable `T-RFLP` without zeros. The plots of the fitted ZAGOLLMax, ZAGA and ZAIG density functions and the histogram are displayed in Figure 4.9a. Plots of the fitted ZAGOLLMax, ZAGA and ZAIG cumulative and empirical cdfs are given in Figure 4.9b. Based on this marginal analysis, we verified that the proposed ZAGOLLMax model can be a good alternative for this type of data, especially when in the occurrence of bimodality.

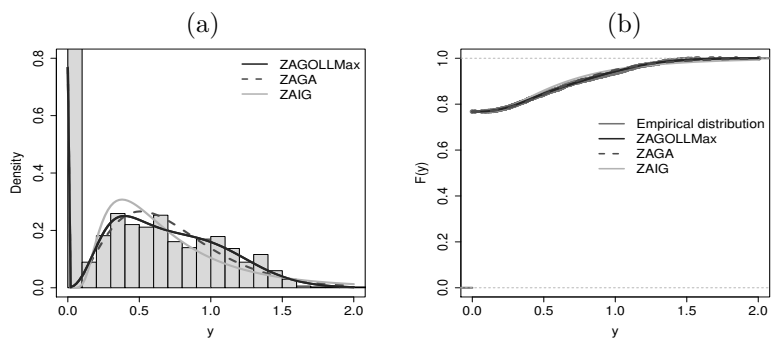

Figure 4.9. Plots for `T-RFLP` variable with the zeros. (a) Histogram with estimated density functions of the GOLLMax, GA and IG models. (b) Empirical distribution with estimated cumulative functions of the GOLLMax, GA and IG models.

4.6.2 The ZAGOLLMax semiparametric regression model

We have the following explanatory variables:

- $x_1 =$ `16S RNAr`; $x_2 =$ `ITS` e $x_3 = phoD$;
- $\underbrace{\text{Block 1 - Block 2 - Block 3 - Block 4}}_{\text{Blocks}} \quad \Rightarrow \quad \underbrace{x_4 \quad x_5 \quad x6}_{\text{3 dummy variables}}$
- $\underbrace{\text{Treat 1 - Treat 2 - Treat 3 - Treat 4 - Treat 5 - Treat 6 - Treat 7}}_{\text{Treatments}} \quad \Rightarrow \quad \underbrace{x_7 \quad x_8 \quad x_9 \quad x_{10} \quad x_{11} \quad x_{12}}_{\text{6 dummy variables}}$
- $\underbrace{\text{Period 1 - Period 2}}_{\text{Period}} \quad \Rightarrow \quad \underbrace{x_{13}}_{\text{1 dummy variable}}$

Then, all variables involved in the study are:

- y_i: response variable `T-RFLP`;
- x_{i1}: `16S RNAr`;
- x_{i2}: `ITS`;
- x_{i3}: *phoD*;
- x_{i4}: comparing Block 1 with Block 2;
- x_{i5}: comparing Block 1 with Block 3;
- x_{i6}: comparing Block 1 with Block 4;
- x_{i7}: comparing Treat 1 with Treat 2;
- x_{i8}: comparing Treat 1 with Treat 3;
- x_{i9}: comparing Treat 1 with Treat 4;
- x_{i10}: comparing Treat 1 with Treat 5;
- x_{i11}: comparing Treat 1 with Treat 6;
- x_{i12}: comparing Treat 1 with Treat 7;
- x_{i13}: comparing period 1 with period 2; for $i = 1, \ldots, 3360$.

We now present results from the fit of the regression model (4.7) under two systematic structures

$$\mu_i = \exp\left(\beta_{01} + \sum_{j=1}^{3} h_j(x_{ij}) + \sum_{j=4}^{13} \beta_{j1}\, x_{ij}\right)$$

and

$$\tau_i = \frac{\exp\left(\beta_{04} + \sum_{j=1}^{3} h_j(x_{ij}) + \sum_{j=4}^{13} \beta_{j4}\, x_{ij}\right)}{1 + \exp\left(\beta_{04} + \sum_{j=1}^{3} h_j(x_{ij}) + \sum_{j=4}^{13} \beta_{j4}\, x_{ij}\right)}$$

The values of the GD, AIC and BIC statistics and the effective degrees of freedom fitted ZAGOLLMax, ZAGA and ZAIG semiparametric regression models are listed in Table 4.10. By comparing these figures, we conclude that the ZAGOLLMax semiparametric regression model outperforms the ZAGA and ZAIG models irrespective of the criteria and then the proposed regression model can be used effectively in the analysis of these data.

Table 4.10. Semiparametric regression models and statistics: GD, AIC and df.

Model	GD	AIC	BIC	df
ZAGOLLMax	3949.4	4045.4	4339.2	48.0
ZAGA	3959.6	4053.6	4341.3	47.0
ZAIG	4153.0	4247.0	4534.6	47.0

We consider the 5% level of significance. Table 4.11 lists the MLEs for the ZAGOLLMax, ZAGA and ZAIG regression models. In this situation, we have that the fixed reference levels are: treatment 1, block 1 and period 1.

The proposed ZAGOLLMax model presents greater robustness in the results when compared to the ZAIG and ZAGA models, in agreement with the descriptive analysis. Note that the proposed model parameter μ is related to the dispersion. Thus, we note that only treatment 2 and treatment 3 differs from treatment 1 at the dispersion parameter and treatment 2, treatment 3, treatment 4 and treatment 6 differ from treatment 1 at the zero inflated parameter. The coefficients of the nonlinear terms are presented, however such values are not interpretable. Our interest is in verifying the forms or trends that the variables x_1, x_2 and x_3 present.

Table 4.11. MLEs of the parameters of the ZAGOLLMax, ZAGA and ZAIG models for microbiological data, the corresponding SEs and p-value.

	ZAGOLLMax			ZAGA			ZAIG		
Parameters	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
β_{01}	-4.667	2.706	0.084	-5.526	2.867	**0.054**	-5.178	3.819	0.175
cs(x_1)	0.209	0.290	0.471	0.335	0.319	0.293	0.352	0.419	0.400
cs(x_2)	0.294	0.091	**0.001**	0.419	0.101	**0.000**	0.406	0.132	**0.002**
cs(x_3)	0.004	0.125	0.973	-0.074	0.146	0.612	-0.141	0.178	0.426
β_{41}	0.079	0.059	0.176	0.022	0.064	0.721	0.059	0.088	0.498
β_{51}	-0.012	0.058	0.831	-0.080	0.062	0.196	-0.070	0.082	0.394
β_{61}	-0.130	0.058	**0.025**	-0.218	0.062	**0.000**	-0.200	0.082	**0.015**
β_{71}	0.134	0.071	**0.059**	0.131	0.074	0.080	0.119	0.099	0.230
β_{81}	0.129	0.063	**0.042**	0.115	0.070	0.103	0.095	0.090	0.292
β_{91}	0.104	0.059	0.076	0.131	0.066	**0.048**	0.132	0.085	0.124
β_{101}	-0.086	0.058	0.138	-0.151	0.066	**0.022**	0.180	0.079	**0.022**
β_{111}	0.117	0.069	0.092	0.113	0.078	0.151	0.095	0.099	0.337
β_{121}	0.023	0.055	0.669	0.002	0.062	0.974	-0.005	0.077	0.945
β_{131}	-0.044	0.052	0.393	-0.047	0.059	0.423	-0.039	0.078	0.613
log(σ)	0.874	0.227	-	-0.277	0.025	-	-0.679	0.024	-
log(ν)	-0.729	0.138	-	-	-	-	-	-	-
β_{04}	-5.146	6.338	0.416	-5.146	6.338	0.416	-5.146	6.338	0.416
$cs(x_1)$	0.205	0.695	0.767	0.205	0.695	0.767	0.205	0.695	0.767
$cs(x_2)$	0.912	0.233	**<0.001**	0.912	0.233	**<0.001**	0.912	0.233	**<0.001**
$cs(x_3)$	-0.180	0.319	0.571	-0.180	0.319	0.571	-0.180	0.319	0.571
β_{44}	0.164	0.138	0.236	0.164	0.138	0.236	0.164	0.138	0.236
β_{54}	0.187	0.136	0.169	0.187	0.136	0.169	0.187	0.136	0.169
β_{64}	-0.517	0.135	**0.0001**	-0.517	0.135	**0.0001**	-0.517	0.135	**0.0001**
β_{74}	0.528	0.172	**0.002**	0.528	0.172	**0.002**	0.528	0.172	**0.002**
β_{84}	0.558	0.161	**0.0005**	0.558	0.161	**0.0005**	0.558	0.161	**0.0005**
β_{94}	0.493	0.153	**0.001**	0.493	0.153	**0.001**	0.493	0.153	**0.001**
β_{104}	-0.188	0.157	0.230	-0.188	0.157	0.230	-0.188	0.157	0.230
β_{114}	0.693	0.177	**<0.001**	0.693	0.177	**<0.001**	0.693	0.177	**<0.001**
β_{124}	0.098	0.147	0.501	0.098	0.147	0.501	0.098	0.147	0.501
β_{134}	-0.133	0.128	0.300	-0.133	0.128	0.300	-0.133	0.128	0.300

4.6.3 Model checking

In Figure 4.10, we perform the residual analysis by plotting the quantile residuals $\hat{rq}_i$ (see Section 4.4) against the index of observations for the fitted ZAGOMAx (Figure 4.10a), ZAGA (Figure 4.10b) and ZAIG (Figure 4.10c) semiparametric regression models in the interval $[-3, 3]$. It is found that the ZAGA model has the highest number of points outside the range $[-3, 3]$.

Figures 4.11, 4.12 and 4.13 give the plots of the quantile residuals by density, qq-plot and worm plot for the ZAGOLLMax, ZAGA and ZAIG models, respectively, to detect possible outlying observations as well as departures from the assumptions of semiparametric regression models. Note that Figures 4.12b, c and 4.13b, c show that the quantile residuals show normality deviations and extrapolation with respect to the confidence bands in the worm plots. Thus, according to Figures 4.10a and 4.11 and Table 4.10 the ZAGOLLMAx model is an alternative to analyze these data under study.

For the ZAGOLLMax semiparametric regression model Figures 4.14 displays the quantile residuals and the envelope. In Figure 4.14a the envelope for the complete model is presented considering the

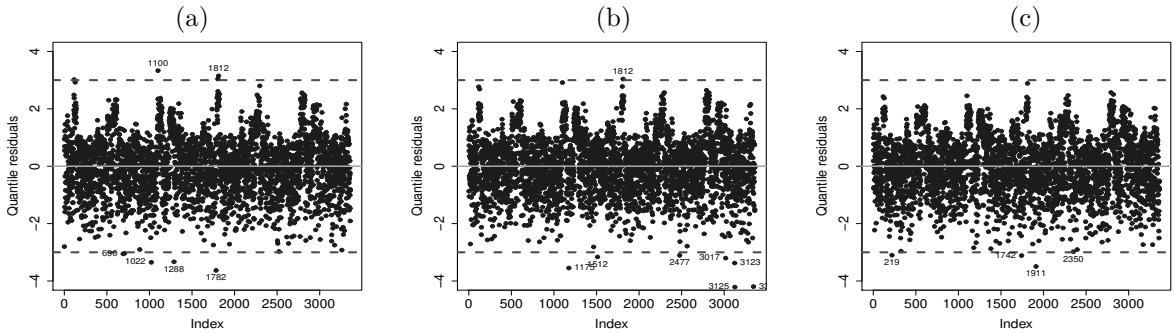

Figure 4.10. Plots of the quantile residuals against the index for microbiological data. (a) For the ZAGOLLMax semiparametric regression model. (b) For the ZAGA semiparametric regression model. (c) For the ZAIG semiparametric regression model.

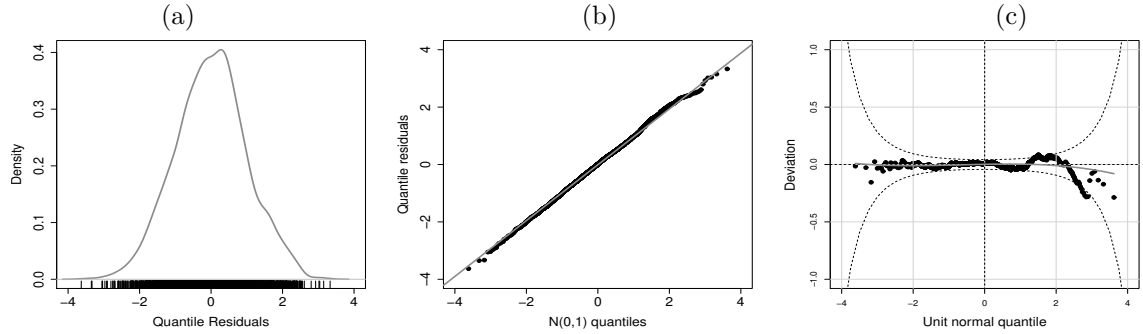

Figure 4.11. Plots of the fitted ZAGOLLMax regression model for microbiological data. (a) Density of the quantile residuals. (b) Q-Q plot for quantile residuals. (c) Worm plot for quantile residuals.

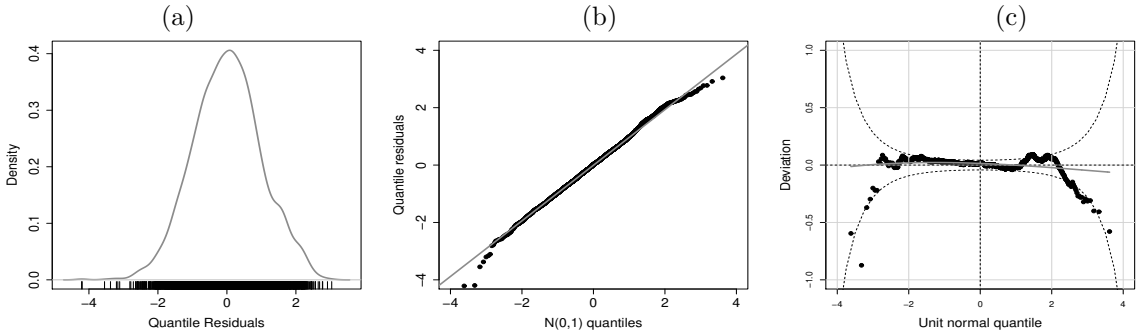

Figure 4.12. Plots of the fitted ZAGA regression model for microbiological data. (a) Density of the quantile residuals. (b) Q-Q plot for quantile residuals. (c) Worm plot for quantile residuals.

systematic regression structure in μ and τ. It is verified that the model is capable of reproducing by simulation the real data. In Figure 4.14b we show for the μ continuous component the simulated envelope where values are generated by the GOLLMax. Figure 4.14c shows the simulated envelope for the discrete component τ, and the values are simulated by the binomial distribution (see Section 4.4). In the three situations considered, it is verified that the proposed model is capable of encompassing practically all points within the confidence bands. This provides indications of the absence of discrepant observations

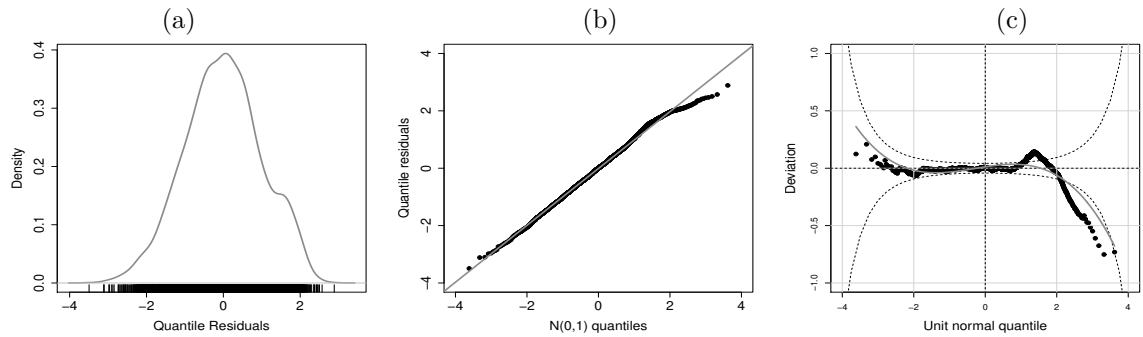

Figure 4.13. Plots of the fitted ZAIG regression model for microbiological data. (a) Density of the quantile residuals. (b) Q-Q plot for quantile residuals. (c) Worm plot for quantile residuals.

and that the ZAGOLLMax model is adequate for this analysis.

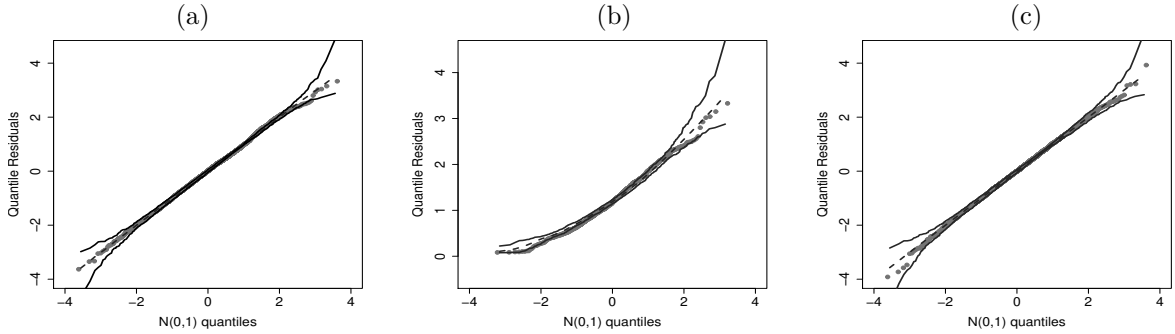

Figure 4.14. Normal probability plot for the quantile residual with envelope based on the fitted ZAGOLLMax regression model to microbiological data. (a) For the complete model with regression structure in μ and τ. (b) For the continuous component with regression structure in μ. (c) For the discrete component with regression structure in τ.

1. **Interpretations of the systematic structure for μ.**

   As explained in the previous section, treatment 1 (control), block 1 and period 1 (6 months) were taken as reference cases. However, the main objective (from the microbiological and agricultural standpoints) is to make comparisons between the treatments, the reason for our discussion in this line.

   - As can be noted in Table4.11, irrespective of the cultivation period, treatments 2 and 3 (compost with and without inoculation) were the only ones that differed (p-value$<$ 0.05) from treatment 1 (control) in relation to the response variable y (`T-RFLP`).
   - If considering a p-value$<$ 0.10, treatments 4 and 6 also differ from treatment 1. This is interesting, since these treatments presented higher productivity than treatment 1, with the greatest yield being observed in the presence of inoculation with phosphate solubilizing bacteria (`PSB`), which was 10% higher (equivalent to approximately 15 tons.ha^{-1}).
   - With this, according to the data obtained by the `T-RFLP` technique, irrespective of the growing period, analysis with the ZAGOLLMax model produces the same productive response found at the end of the cultivation period (one year after planting).

- The differences found in relation to treatment 1 reinforce the efficacy of applying the compost, even without enrichment with phosphate sources, especially the presence of `BSF`.
- Some authors have found increased yield of sugarcane with the application of filter cake associated with phosphate rocks, but have not considered the microbiological aspects (Teles *et al.* 2017 and Soltangheisi *et al.* 2019).
- The data on productivity (to be published) corroborate the findings with application of the ZAGOLLMax model, making it applicable to the dataset under analysis.
- For the covariables, we observed, regarding the shape presented from the smoothing curve, that there was no substantial alteration of x_1 (`16S RNAr`) (Figure 4.15a), while for x_2 (`ITS`) there was a tendency for increase (Figure 4.15b). For the covariable x_3 (*phoD*), there was a tendency to increase up to 6.6, followed by a decrease in the shape of the smoothing curve (Figure 4.15c).

2. **Interpretations of the systematic structure for τ.**

   Analysis of the proportion of zeros shown in Table 4.11 (lower part) indicates that:

   - The data from treatments 2, 3, 4 and 6 have a greater proportion of zeros than those of treatment 1.
   - In treatment 4, there was enrichment of the organic compost with apatite (an igneous rock, harder to became available to plants) and inoculation with `PSB`.
   - In treatment 6, there was enrichment of the compost with phosphorite (of sedimentary origin, easier to become available) without the inoculation with `PSB`.
   - The differences in the proportions of zeros found in the data from these treatments denote differences in the fungal community structure. This higher proportion of zeros observed from treatments 2,3,4 and 6 leads to the interpretation that these treatments share absence of determined `T-RFs`.
   - Nevertheless, this absence of `T-RFs` might be contributing to increase the productivity, since these treatments showed higher productivity than treatment 1.
   - In the covariables, we observed the same behavior as in μ, but there was a rising trend for the variable x_2 (`ITS`) and for $x_3 > 0.7$ (*phoD*) there was an approximately linear decline.

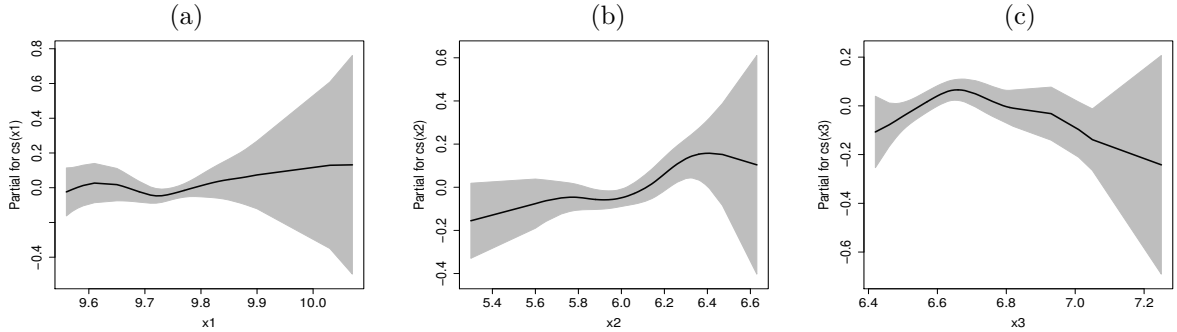

Figure 4.15. Fitted terms for ZAGOLLMax regression model with regression structure in μ parameter. (a) For x_1. (b) For x_2. (c) For x_3.

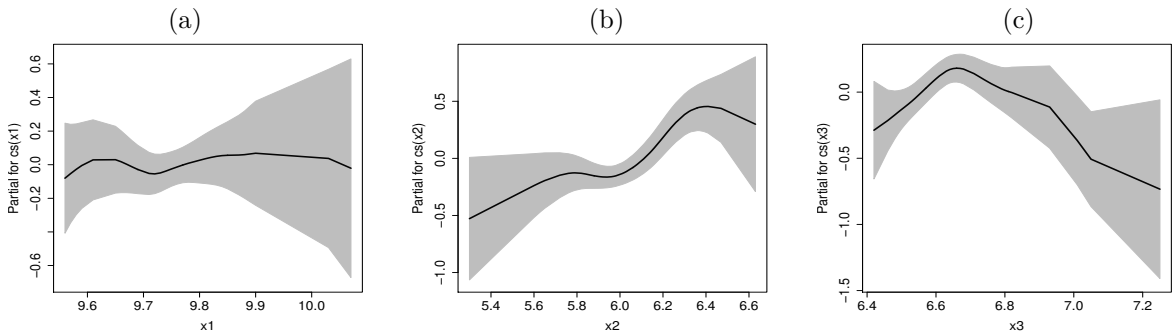

Figure 4.16. Fitted terms for ZAGOLLMax regression model with regression structure in τ parameter. (a) For x_1. (b) For x_2. (c) For x_3.

4.7 Concluding Remarks

We define a new *zero adjusted generalized odd log-logistic Maxwell* (ZAGOLLM) semiparametric regression to analyze data in presence of bimodality, heteroskedasticity, zero-inflation and nonlinear effects in covariables. We discuss some inferential issues related to this regression and perform some simulations. We illustrate the potentiality of the new regression by means of microbiological data which includes descriptive and marginal analysis, model checking and interpretations of its systematic components. In conclusion, the proposed semiparametric regression is extremely effective in demonstrating all the biological and productive effects observed in the field. The organic compost, when inoculated with PSB, was sufficient to evidence 10% increase in sugarcane productivity even in the absence of rock phosphate.

References

Andreote, F. D. and e Silva, M. D. C. P. (2017). Microbial communities associated with plants: learning from nature to apply it in agriculture. *Current opinion in microbiology*, **37**, 29-34.

Aarts, R. M. (2000). Lauricella functions, www.mathworld.wolfram.com/LauricellaFunctions.html. From MathWorld - A Wolfram Web Resource, created by Eric W. Weisstein.

Buuren, S. V. and Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, **20**, 1259-1277.

Cordeiro, G. M., Alizadeh, M., Ozel, G., Hosseini, B., Ortega, E. M. M. and Altun, E. (2017). The generalized odd log-logistic family of distributions: properties, regression models and applications. *Journal of Statistical Computation and Simulation*, **87**, 908-932.

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

Estrada-Bonilla, G. A., Lopes, C. M., Durrer, A., Alves, P. R., Passaglia, N. and Cardoso, E. J. (2017). Effect of phosphate-solubilizing bacteria on phosphorus dynamics and the bacterial community during composting of sugarcane industry waste. *Systematic and applied microbiology*, **40**, 308-313.

Gardes, M. and Bruns, T. D. (1993). ITS primers with enhanced specificity for basidiomycetes?application to the identification of mycorrhizae and rusts. *Molecular ecology*, **2**, 113-118.

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

Gurdeep, K. A. U. R. and Reddy, M. S. (2015). Effects of phosphate-solubilizing bacteria, rock phosphate and chemical fertilizers on maize-wheat cropping cycle and economics. *Pedosphere*, **25**, 428-437.

Hashimoto, E. M., Ortega, E. M. M., Cordeiro, G. M., Cancho, V. G. and Klauberg, C. (2019). Zero-spiked regression models generated by gamma random variables with application in the resin oil production. *Journal of Statistical Computation and Simulation*, **89**, 52-70.

Heller, G., Stasinopoulos, M. and Rigby, B. (2006, July). The zero-adjusted inverse Gaussian distribution as a model for insurance claims. *In Proceedings of the 21th International Workshop on Statistical Modelling*. sn.

Lambais, M. R., de Carvalho, J. and de Campos Büll, R. (2005). Diversidade Microbiana nos Solos: Definindo novos paradigmas. *Tópicos em Ciência do Solo*, Viçosa, **4**, 42-84.

Liesack, W., and Dunfield, P. F. (2004). *T-rflp analysis*. In Environmental Microbiology (pp. 23-37). Humana Press.

Muyzer, G., De Waal, E. C. and Uitterlinden, A. G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.*, **59**, 695-700.

Ospina, R., and Ferrari, S.L.P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics and Data Analysis*, **56**, 16091623.

Ramires, T. G., Hens, N., Cordeiro, G. M. and Ortega, E. M. (2018). Estimating nonlinear effects in the presence of cure fraction using a semi-parametric regression model. *Computational Statistics*, **33**, 709-730.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric regression*. Cambridge university press.

Sakurai, M., Wasaki, J., Tomizawa, Y., Shinano, T. and Osaki, M. (2008). Analysis of bacterial communities on alkaline phosphatase genes in soil supplied with organic matter. *Soil science and plant nutrition*, **54**, 62-71.

Soltangheisi, A., Santos, V. R. D., Franco, H. C. J., Kolln, O., Vitti, A. C., Dias, C. T. D. S., ... and Pavinato, P. S. (2019). Phosphate Sources and Filter Cake Amendment Affecting Sugarcane Yield and Soil Phosphorus Fractions. *Revista Brasileira de Ciência do Solo*, **43**.

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

Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., Voudouris V. and De Bastiani, F. (2017), *Flexible Regression and Smoothing: Using GAMLSS in R*, Chapman and Hall/CRC, New York.

Teles, A. P. B., Rodrigues, M., Bejarano Herrera, W. F., Soltangheisi, A., Sartor, L. R., Withers, P. J. A. and Pavinato, P. S. (2017). Do cover crops change the lability of phosphorus in a clayey subtropical soil under different phosphate fertilizers?. *Soil use and management*, **33**, 34-44.

Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J. and Stasinopoulos, D. (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, **39**, 1279-1293.

White, T. J., Bruns, T., Lee, S., Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis, M., Gelfand, D., Sninsky, D., White, T., eds. **PCR protocols: a guide to methods and applications**. Orlando, Academic Press, p. 315-322.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

5 THE GENERALIZED ODD LOG-LOGISTIC MAXWELL CURE RATE SEMIPARAMETRIC REGRESSION MODELS APPLIED TO PROSTATE CANCER DATA

Abstract: In this paper we propose a new semiparametric regression model with generalized odd log-logistic Maxwell errors to model possible presence of long-term survivors in the data using the cubic splines basis for nonlinear effects. The models attempt to simultaneously estimate the effects of covariates on the acceleration/deceleration of the timing of a given event and the surviving fraction, that is, the proportion of the population for which the event never occurs. We consider penalized quasi likelihood estimators for the fixed and random parameters of the model. Finally, we analyze a real data set for localized prostate cancer patients after open radical prostatectomy.
Keywords: Cubic spline; Censored data; Maxwell distribution; Penalized likelihood; Prostate cancer.

5.1 Introduction

In addition to choosing the model to be used in the study, one of the basic assumptions of the survival models is that all individuals will present the event of interest since they are followed for a sufficiently long period of time. However, in some situations, not all individuals will be susceptible to the event of interest for as long as the follow-up time is, these individuals are called immune or cured (Maller and Zhou, 1996). Long-life models or models with a curing fraction adjust data with these characteristics. The best known long-term models are: the standard mixing model, initially introduced by Boag (1949) and developed by Berkson and Gage (1952), and the promotion time model, proposed by Yakovlev *et al.* (1993). The standard mixture model divides the population into two groups and consists of a mixture of two distributions for these two groups. On the other hand, the promotion time model involves a competitive risk structure in which n factors compete for the occurrence of the event of interest. Rodrigues *et al.* (2009) proposed a unified long-term model whose particular cases are the standard mix model and the promotion time model. Ortega *et al.* (2009) presented the generalized log-gamma regression model with fraction of cure, including as special cases the exponential regression, Weibull and log-normal models with fraction of cure. Martinez *et al.* (2013) use a mixture model and a non-mixture model based on the modified Weibull distribution generalized in gastric cancer data and the results are obtained by Bayesian inference.

Cure rate models have been used to model time-to-event data for various types of cancer, including breast cancer, leukemia, melanoma and prostate cancer. Recently, Ortega *et al.* (2012, 2013) present studies and analyzes related to a database on prostate cancer. Such a database is previously studied and provided by Kattan *et al.* (1999) and Stephenson *et al.* (2005). The database studied by these authors consists of the random response variable given by the number of months without detectable disease after prostatectomy the main objective is to investigate a possible relation of the responser variable and other explanatory variables. This data set basically has three problems:

1. In Figure 5.1a, we can see that the empirical failure rate function has the form of descending, increasing and descending.

2. In Figure 5.1b, we can see that there is a possibility of a proportion of individuals cured.

3. In Figure 5.2a and Figure 5.2b , presence of heterogeneity and nonlinear behavior of the covariable ($x_1 = PSA$) and the response variable time (months).

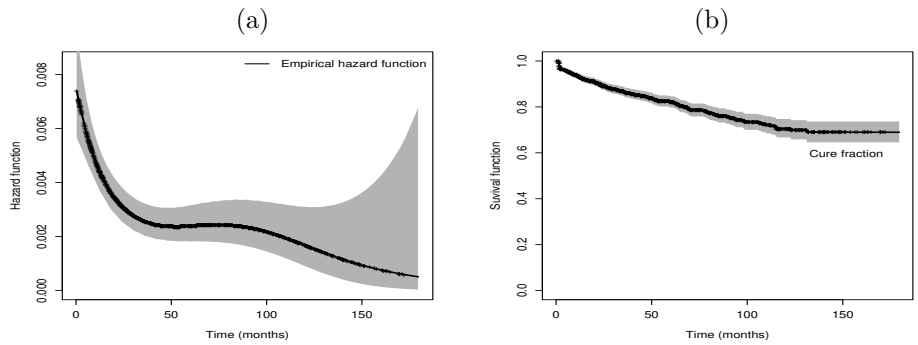

Figure 5.1. (a) Empirical survival function. (b) Empirical hazard function.

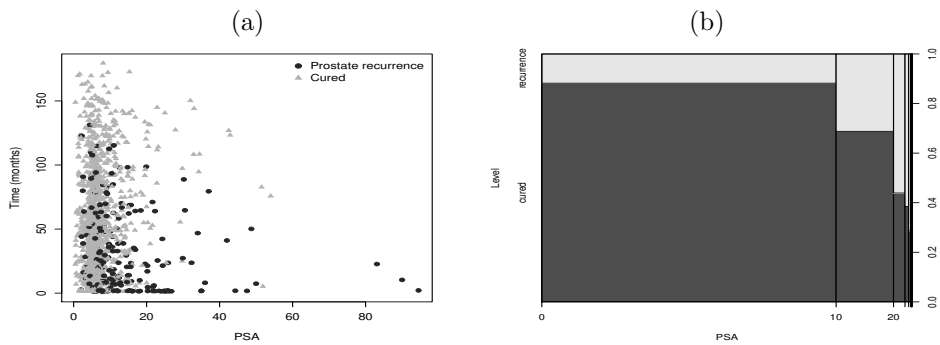

Figure 5.2. (a) Scatter plots between the variable ($x_1 = PSA$) versus the response variable time (months). (b) Spine plot for variable ($x_1 = PSA$) versus proportion of cured and uncured.

So in this research we are proposing a solution for each problem presented in items (1), (2) and (3).

1. Solution for (1). As the usual distributions do not model this type of failure rate function, we are proposing the distribution called generalized odd log-logistic Maxewell distribution to model different types of failure rate function forms. Note that we are also introducing a new distribution little known in the area of survival analysis which is the Maxwell distribution.

2. Solution for (2). To model the proportion of cured we will combine the mixing model with the GOLLMax distribution, including the regression structure both in the time of failure as well as in the proportion of cured. Perhaps the most popular type of cure rate models is the mixture model introduced by Berkson and Gage (1958) and Maller and Zhou (1996).

3. Solution for (3). To model nonlinear effects we are going to propose a GOLLMax mixture semi-parametric regression model.

Based on these 3 solutions, the following research proposes the *generalized odd log-logistic Maxwell mixture* (GOLLMaxM) distribution from the GOLLMax. In this chapter, we propose a new GOLLMaxM semiparametric regression model with a cure rate to analyze survival data of long-term survivors to solve the problems above.

The rest of the chapter proceeds as follows. In Section 5.2, we formulate the model and derive the time distribution for the entire population. In Section 5.3, we present the he generalized odd log-

logsitic Maxwell mixtures semi-parametric regression model and inference based on penalized maximum likelihood is developed A simulation study is presented in Section 5.4 in order to verify some statistical properties of the penalized maximum likelihood estimators. An application of the proposed model to fit a prostate cancer data set is reported in Section 5.5, which illustrates the potential of our methodology. Finally, some conclusions are given in Section 5.6.

5.2 Model formulation

A random variable T has the GOLLMax distribution if its cumulative distribution function (cdf) and probability density function (pdf) are (for $t > 0$)

$$F(t;\mu,\sigma,\nu) = \frac{\gamma_1^{\nu\,\sigma}(3/2, t^2/\mu^2)}{\gamma_1^{\nu\sigma}(3/2, t^2/\mu^2) + [1-\gamma_1^{\sigma}(3/2, t^2/\mu^2)]^{\nu}} \tag{5.1}$$

and

$$f(t;\mu,\sigma,\nu) = \frac{4\,\nu\,\sigma\,t^2}{\sqrt{\pi}\,\mu^3}\,\exp\left(-\frac{t^2}{\mu^2}\right)\frac{\gamma_1^{\nu\sigma-1}(3/2, t^2/\mu^2)\left[1-\gamma_1^{\sigma}(3/2, t^2/\mu^2)\right]^{\nu-1}}{\{\gamma_1^{\nu\sigma}(3/2, t^2/\mu^2) + [1-\gamma_1^{\sigma}(3/2, t^2/\mu^2)]^{\nu}\}^2}, \tag{5.2}$$

respectively, where $\nu > 0$ and $\sigma > 0$ are two extra shape parameters and $\mu > 0$ is scale parameter. The $\gamma_1(p,u) = \gamma(p,u)/\Gamma(p)$ is the incomplete gamma function ratio, $\gamma(p,u) = \int_0^u w^{p-1}\mathrm{e}^{-\mathrm{w}}\mathrm{dw}$ is the incomplete gamma function and $\Gamma(\cdot)$ is the gamma function. Henceforth, if Y is a random variable with cdf (5.1), we write $T \sim \mathrm{GOLLMax}(\mu,\nu,\sigma)$. The hazard rate function (hrf) of Y is given by $h(t;\mu,\sigma,\nu) = f(t;\mu,\sigma,\nu)/[1-F(t;\mu,\sigma,\nu)]$, where $1-F(t;\mu,\sigma,\nu)$ is survival function.

To formulate the generalized odd log-logistic Maxwell mixture model (GOLLMaxM), we considered that the studied population is a mixture of susceptible (uncured) individuals, who may experience the event of interest, and non-susceptible (cured) individuals, who will experience it (Maller and Zhou, 1996). This approach allows to estimate simultaneously whether the event of interest will occur, which is called incidence, and when it will occur, given that it can occur, which is called latency. Let N_i (for $i = 1,\ldots,n$) be the indicator denoting that the ith individual is susceptible ($N_i = 1$) or non-susceptible ($N_i = 0$), i.e., the population is classified in two sub-populations so that an individual either is cured with probability $0 < \tau < 1$, or has a proper survival function $S(t)$ with probability $(1-\tau)$. The mixture model (MM) can be expressed by

$$S_{pop}(t_i) = \tau + \big(1-\tau\big)S(t_i|N_i = 1), \tag{5.3}$$

where $S_{pop}(t_i)$ is the unconditional survival function of t_i for the entire population, $S(t_i|N_i = 1)$ is the survival function for susceptible individuals and $\tau = P(N_i = 0)$ is the probability of cure of an individual. The population survival function, denoted by $S_{pop}(t_i)$, for GOLLMaxM model is given by,

$$S_{pop}(t) = \tau + \frac{\big(1-\tau\big)[1-\gamma_1^{\sigma}(3/2, t^2/\mu^2)]^{\nu}}{\gamma_1^{\nu\sigma}(3/2, t^2/\mu^2) + [1-\gamma_1^{\sigma}(3/2, t^2/\mu^2)]^{\nu}}, \tag{5.4}$$

where $S_{pop}(t)$ is the unconditional survival function of T for the entire population, $S(t|N = 1)$ the survival function for susceptible individuals and $\tau = P(N = 0)$ is the probability of cure variation, $(0 < \tau < 1)$. The pdf corresponding to (5.4) is given by

$$f_{pop}(t) = \frac{(1-\tau)\,4\,\nu\,\sigma\,t^2}{\sqrt{\pi}\,\mu^3}\,\exp\left(-\frac{t^2}{\mu^2}\right)\frac{\gamma_1^{\nu\sigma-1}(3/2, t^2/\mu^2)\left[1-\gamma_1^{\sigma}(3/2, t^2/\mu^2)\right]^{\nu-1}}{\{\gamma_1^{\nu\sigma}(3/2, t^2/\mu^2) + [1-\gamma_1^{\sigma}(3/2, t^2/\mu^2)]^{\nu}\}^2},. \tag{5.5}$$

The hazard rate function (hrf) of the GOLLMaxM model is given by $h_{pop}(t) = f_{pop}(t)/S_{pop}(t)$. A random variable having density (5.5) is denoted by $T \sim GOLLMaxM(\mu,\sigma,\nu,\tau)$. The GOLLMaxM model contains as special cases the following distributions:

- For $\sigma = 1$, it leads to the (new) odd log-logistic Maxwell mixture (OLLMaxM) model.
- For $\nu = 1$, it gives the (new) exponentiated Maxwell mixture (EMaxM) model .
- The (new) Maxwell mixture (MM) model is as a basic exemplar when $\sigma = \nu = 1$.

Plots of the GOLLMaxM survival and hazard functions for selected parameter values are displayed in Figures 5.3a and 5.3b, respectively.

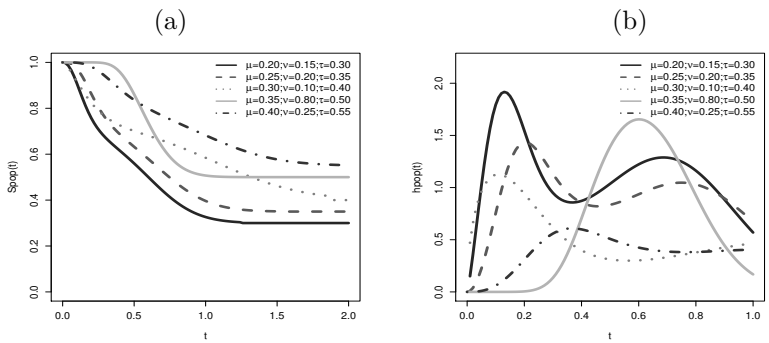

Figure 5.3. (a) The GOLLMaxM survival function for fixed $\sigma = 5.0$. (b) The GOLLMaxM hazard function for fixed $\sigma = 5.0$.

5.3 The GOLLMaxM regression model with cure fraction

In order to introduce a regression structure in the class of models (5.4), we assume that both parameters μ_i and τ_i vary across observations through regression structures which are parameterized as

$$\mu_i = \mu_i(\boldsymbol{\beta}_1), \quad \tau_i = \tau_i(\boldsymbol{\beta}_2),$$

where $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1q_1})^\top$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2q_2})^\top$. The usual systematic component for the scala parameter is $\mu_i = \exp(\mathbf{x}_{1i}^\top \boldsymbol{\beta}_1)$, where $\mathbf{x}_{1i}^\top = (x_{1i1}, \ldots, x_{1iq_1})$ is a vector of known explanatory variables, i.e. $\boldsymbol{\mu} = \exp(\boldsymbol{X_1}\boldsymbol{\beta_1})$, with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ and $\boldsymbol{X_1} = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n})^\top$ is a specified $n \times q_1$ matrix of full rank and $q_1 < n$.

Analogously, we consider for the cured proportion parameter the systematic component

$$\tau_i = \frac{\exp(\mathbf{x}_{2i}^\top \boldsymbol{\beta}_2)}{1 + \exp(\mathbf{x}_{2i}^\top \boldsymbol{\beta}_2)}, \quad \text{where} \quad \boldsymbol{x}_{2i}^\top = (x_{2i1}, \ldots, x_{2iq_2}),$$

is a vector of known explanatory variables; i.e. the linear structure, then have $\boldsymbol{\tau} = \frac{\exp(\boldsymbol{X_2}\boldsymbol{\beta_2})}{1+\exp(\boldsymbol{X_2}\boldsymbol{\beta_2})}$, where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^\top$ and $\boldsymbol{X_2} = (\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n})^\top$ is a specified $n \times q_2$ matrix of full rank and $q_2 < n$. The dispersion covariates in $\boldsymbol{X_2}$ are commonly, but not necessary, a subset of the regression covariates in $\boldsymbol{X_1}$. It is assumed that $\boldsymbol{\beta_1}$ is functionally independent of $\boldsymbol{\beta_2}$. The identifiability between the parameters in the cure fraction and those in the latency distribution for the mixture model has been discussed by Li *et al.* (2001). The mixture model is not identifiable when cure fraction $\tau(\boldsymbol{\beta}_2)$ is a constant τ, but is identifiable when $\tau(\boldsymbol{\beta}_2)$ is modeled by a logistic regression with non-constant covariates $\mathbf{x}_2$ (Li *et al.* 2001). So, it is necessary to include some covariates in the cure fraction to ensure identifiability. Note that τ_i is the probability of cure variation from individual to individual.

5.3.1 Extension semi-parametric

In the surveys that consider regression models, the structure of continuous covariates is added in linear models in the parameters both in the proportion of individuals cured, as well as in the time of failure, although this relationship is not always true. That is, in some cases we can have covariates that have a non-linear relationship, so to capture the non-linear effects of these covariates, it is necessary to adopt non-linear functions.

Let $\mathbf{x}_{3i}^{\top} = (x_{3i1}, \ldots, x_{3iq_3})$ be the vector of covariates that has a nonlinear form with the response variable, we can define semi-parametric structures for the elements of the vector $\boldsymbol{\theta}$ using appropriate link functions as

$$\mu_i = \exp\left(\mathbf{x}_{1i}^{\top}\boldsymbol{\beta}_1 + \sum_{j=1}^{J} h_j(\mathbf{x}_{j3i})\right) \qquad \tau_i = \frac{\exp\left(\mathbf{x}_{2i}^{\top}\boldsymbol{\beta}_2 + \sum_{j=1}^{J} h_j(\mathbf{x}_{j3i})\right)}{1 + \exp\left(\mathbf{x}_{2i}^{\top}\boldsymbol{\beta}_2 + \sum_{j=1}^{J} h_j(\mathbf{x}_{j3i})\right)} \tag{5.6}$$

where $h_j(.)$ are smooth functions of the covariates $\mathbf{x}_{3i}$ for $j = 1, \ldots, J$ and $i = 1, \ldots, n$.

The model (5.6) will be referred to as the GOLLMaxM semi-parametric regression model with long-term survivors in competitive-risk structure. The most important of this regression model defines the parameters depending on $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$ and $\mathbf{x}_{3i}$. In this paper, we only use the cubic splines as smooth functions $h_j(.)$. The GOLLMaxM regression model (5.4) opens new possibilities for fitting many different types of data, since the GOLLMax distribution is much more flexible then the Maxwell distribution and hence data with monotone and nonmonote hazard rate functions can be analyzed using the proposed regression model.

Suppose we have data in the form $(t_i, \mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i})$, $i = 1, \ldots, n$ where t_i denotes the observed time for the ith subject, i.e, $t_i = \min\{T_i, C_i\}$, T_i is the lifetime for the ith individual and C_i is the censoring time for the ith individual, $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ is covariate vectors. With this assumption we have, that the contribution of an individual that failed at t_i to the likelihood function is given by

$$\frac{(1-\tau_i)4\,\nu\,\sigma\,t_i^2}{\sqrt{\pi}\,\mu_i^3}\,\exp\left(-\frac{t_i^2}{\mu_i^2}\right)\frac{\gamma_1^{\nu\sigma-1}(3/2, t_i^2/\mu_i^2)\left[1-\gamma_1^{\sigma}(3/2, t_i^2/\mu_i^2)\right]^{\nu-1}}{\left\{\gamma_1^{\nu\sigma}(3/2, t_i^2/\mu_i^2) + \left[1-\gamma_1^{\sigma}(3/2, t_i^2/\mu_i^2)\right]^{\nu}\right\}^2} \tag{5.7}$$

and the contribution of an individual that is at risk at t_i is

$$\tau_i + \frac{\left(1-\tau_i\right)\left[1-\gamma_1^{\sigma}(3/2, t_i^2/\mu_i^2)\right]^{\nu}}{\gamma_1^{\nu\sigma}(3/2, t_i^2/\mu_i^2) + \left[1-\gamma_1^{\sigma}(3/2, t_i^2/\mu_i^2)\right]^{\nu}}. \tag{5.8}$$

For the semiparametric model (5.6), the fixed and random effects $\boldsymbol{\theta} = (\sigma, \nu, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\boldsymbol{\eta}$, respectively, are estimated by maximizing the penalized log-likelihood function

$$\begin{aligned} l(\boldsymbol{\omega}) \;=\; & r\log\left(\frac{4\,\nu\,\sigma}{\sqrt{\tau}}\right) + \sum_{i\in F}\log(1-\tau_i) + \sum_{i\in F}\log\left(\frac{t_i^2}{\mu_i^3}\right) - \sum_{i\in F}\left(\frac{t_i}{\mu_i}\right)^2 + \\ & (\nu\,\sigma-1)\sum_{i\in F}\log\left[\gamma_1(3/2, t_i^2/\mu_i^2\right] + (\nu-1)\sum_{i\in F}\log\left[1-\gamma_1^{\sigma}(3/2, t_i^2/\mu_i^2)\right] - \\ & 2\sum_{i\in F}\log\left\{\gamma_1^{\nu\,\sigma}(3/2, t_i^2/\mu_i^2) + \left[1-\gamma_1^{\sigma}(3/2, t_i^2/\mu_i^2)\right]^{\nu}\right\} \\ & + \sum_{i\in C}\log\left\{\tau_i + \frac{\left(1-\tau_i\right)\left[1-\gamma_1^{\sigma}(3/2, t_i^2/\mu_i^2)\right]^{\nu}}{\gamma_1^{\nu\sigma}(3/2, t_i^2/\mu_i^2) + \left[1-\gamma_1^{\sigma}(3/2, t_i^2/\mu_i^2)\right]^{\nu}}\right\} - \frac{1}{2}\sum_{j=1}^{J}\lambda_j\boldsymbol{\eta}_j^T\mathbf{P}_j\boldsymbol{\eta}_j^T, \end{aligned} \tag{5.9}$$

where $\boldsymbol{\omega} = (\boldsymbol{\theta}^T, \boldsymbol{\eta}^T)^{\top}$ r is the number of uncensored observations (failures), F and C denote, respectively, that the set of individuals is a lifetime or a censoring time, $\gamma_1(\cdot,\cdot)$ is the incomplete gamma function

ratio, where $\mathbf{P}_j$ is a symmetric matrix that may depend on a vector of smoothing parameters, see for example, Rigby and Stasinopouls (2005). For each smoothing term selected, and any of the parameters of the GOLLMaxM distribution, there is one smoothing parameter λ associated with it. The smoothing parameters can be fixed or estimated from the data. The numerical maximization of the (5.9) can be performed in the `gamlss` and `gamlss.cens` packages in `R`. We use the maximization by `RS` algorithm described by Stasinopoulos and Rigby (2007) and Stasinopoulos et al. (2017). The `cs()` function is used to assign the arguments to make the adjustment via `gamlss`. Thus, in Section 5.5, the `cs(·)` function in regression structures is denoted by $cs(x_{ji}, df^*)$ where x_{ji} is the j-th covariate considering the additive term and df^* are the degrees of freedom related to the additive term. The effective degree of freedom for structure of the regression in μ considering an explanatory variable x is given by $df_\mu = df^* + 2$ where other two additional degrees of freedom are in relation to the linear terms (see, Voudouris *et al.* (2012)). Finally, we have that the total freedom degree of the adjusted model, represented by df, collectively considers the additive terms represented by the $h_j(\cdot)$ functions and the parametric terms, i.e, $df = df_\mu + df_\sigma + df_\nu + df_\tau$, are the degrees of freedom used to model μ, σ, ν and τ, respectively. Thus, in this study we consider in the simulations and application $df^* = 3$ which is the default value of the $cs(\cdot)$ function.

5.3.2 Choosing the best model

For selection of the appropriate distribution, we use the global deviance (GD), $GD = -2l_p(\hat{\boldsymbol{\omega}})$, $l_p(\hat{\boldsymbol{\omega}})$ is the penalized log-likelihood function and the generalized Akaike information criterion (GAIC) defined by $GAIC(k) = GD + k \times df$, where df is the total degrees of freedom of the adjusted model and k is the penalty for each degree of freedom used. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are special cases of the GAIC(k) measure when $k = 2$ and $k = \log(n)$, respectively. We consider the GD, AIC and BIC measures to select the best models.

5.3.3 Residual analysis

The important step in the analysis of a fitted model is to check possible deviations from the model assumptions. In this context, we consider the quantile residuals (Dunn and Smyth, 1996) for the GOLLMax semiparametric regression model have the form

$$\widehat{rq}_i = \Phi^{-1}\left\{\frac{\gamma_1^{\hat{\nu}\,\hat{\sigma}}(3/2, y^2/\hat{\mu}_i^2)}{\gamma_1^{\hat{\nu}\hat{\sigma}}(3/2, y^2/\hat{\mu}_i^2) + \left[1 - \gamma_1^{\hat{\sigma}}(3/2, y^2/\hat{\mu}_i^2)\right]^{\hat{\nu}}}\right\}, \tag{5.10}$$

where $\Phi^{-1}(\cdot)$ is the standard normal qf.

5.4 Simulation study

In this section, we examine the performance of the GOLLMaxM semiparametric regression model by means of a Monte Carlo simulation study under two scenarios. Various simulations are conducted for different sample sizes ($n = 100, 350, 700$) using the `R` software with `gamlss` packages by RS method.

In this study, we considered two scenarios for GOLLMaxM semiparametric regression model with regression structure defined in equation (5.6):

- The first scenario, we consider the following structure for regression model with parameter $\mu_i = \exp\{h_{21}(x_{1i}) + \beta_{21}x_{2i} + \beta_{31}x_{3i}\}$, where $h_{21}(x_{1i}) = \sin(x_{1i})$ with functional shape presented in Figure 5.4a, $\sigma = \exp\{\beta_{02}\}$, $\nu = \exp\{\beta_{03}\}$ and $\tau_i = \text{logit}\{\beta_{04} + \beta_{14}x_{4i}\}$, the values associated

to the coefficients are: $\beta_{21} = 0.20$, $\beta_{31} = -0.35$, $\beta_{03} = 0.35$, $\beta_{04} = 0.85$, $\beta_{04} = -0.95$ and $\beta_{14} = 0.30$, explanatory variables for the scenario are $x_{1i} \sim \text{Uniform}(0, 2.5)$, $x_{2i} \sim \text{Normal}(5, 0.50)$, $x_{3i} \sim \text{Uniform}(0, 1)$ and $x_{4i} \sim \text{Binomial}(1, 0.5)$.

- The second scenario, we consider the idea presented in (Ramirez *et al.* 2018) for nonlinear effects in regression models with long-term survival. Thus, the following structure for regression model with parameter $\mu_i = \exp\{h_{21}(x_{1i}) + \beta_{21}x_{2i} + \beta_{31}x_{3i}\}$, where $h_{21}(x_{1i}) = \sin(x_{1i})$ with functional shape presented in Figure 5.4a, $\sigma = \exp\{\beta_{02}\}$, $\nu = \exp\{\beta_{03}\}$ and structure of the regression model for parameter $\tau_i = \text{logit}\{h_{41}(x_{4i})\}$ with functional shape presented in Figure 5.4b. The explanatory variables for the scenario are $x_{1i} \sim \text{Uniform}(0, 2.5)$, $x_{2i} \sim \text{Normal}(5, 0.50)$, and $x_{3i} \sim \text{Uniform}(0, 1)$. For each level of x_{4i}, it was generated a sample size of length n_i, so that $n = \sum_{i=1}^{10} n_i$. The fixed values of τ, for each value of the x_{4i}, are given in Table 5.1. The associated to the coefficients are: $\beta_{21} = 0.20$, $\beta_{31} = -0.35$, $\beta_{03} = 0.35$ and $\beta_{04} = 0.95$.

Table 5.1. Fixed values of the τ parameter of each level of the x_4 explanatory variable.

τ	0.20	0.25	0.35	0.40	0.45	0.45	0.40	0.35	0.25	0.20
x_4	1	2	3	4	5	6	7	8	9	10

We assume that the a percentage of cured approximately 0.55 for the two scenarios. To generate the random values of the proposed model with of cured proportion, we present a brief script:

i. Calculation τ such that $\tau_i = \text{logit}(\beta_{04} + \beta_{14}x_i)$;

ii. $M_i \sim Bernoulli(\tau_i)$;

iii. If $M_i = 0$, $y_i = \infty$, else $M_i = 1$, y_i =rGOLLMax(1,μ_i,σ_i,ν_i);

iv. Generate censored time by $tc_i \sim \text{Uniforme}(0, \xi)$, where ξ denoted the proportion of censored observations, which $\xi = 18$ for first and second scenario, respectively;

v. The observed time t_i of the i-th individual is $t_i = min(y_i, tc_i)$;

vi. Create a censored indicator vector, δ_i, if $y_i \leq tc_i$ do $\delta_i = 1$, otherwise $\delta_i = 0$.

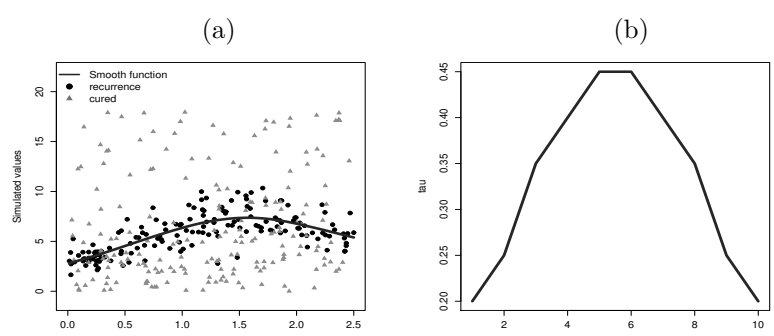

Figure 5.4. Plots of the simulation values. (a) Nonlinear effect for x_1 versus time simulated. (b) Nonlinear effect for x_4 versus cured proportion simulated.

For each of the $1,000$ simulations, the average estimates (AEs), biases and MSEs are calculated. The results are reported in Tables 5.2 and 5.3 for the parametric and semi-parametric models. Based

Table 5.2. The AEs, biases and MSEs for the parametric and semiparametric GOLLMaxM regression models based on 1,000 simulations for scenario 1.

	Semiparametric				Parametric			
n	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.192	-0.008	0.004	β_{21}	0.197	-0.003	0.006
	β_{31}	-0.345	0.005	0.010	β_{31}	-0.339	0.011	0.017
100	β_{02}	0.224	-0.126	0.028	β_{02}	0.117	-0.233	0.060
	β_{03}	0.885	0.035	0.010	β_{03}	0.565	-0.285	0.087
	β_{04}	-1.061	-0.111	2.629	β_{04}	-1.169	-0.219	4.143
	β_{14}	0.406	0.106	2.651	β_{14}	0.506	0.206	3.937
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.196	-0.004	0.001	β_{21}	0.204	0.004	0.001
	β_{31}	-0.325	0.025	0.003	β_{31}	-0.296	0.054	0.007
350	β_{02}	0.230	-0.120	0.019	β_{02}	0.134	-0.216	0.049
	β_{03}	0.827	-0.023	0.003	β_{03}	0.550	-0.300	0.092
	β_{04}	-0.947	0.003	0.044	β_{04}	-0.976	-0.026	0.048
	β_{14}	0.301	0.001	0.086	β_{14}	0.309	0.009	0.091
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.200	0.000	0.000	β_{21}	0.209	0.009	0.001
	β_{31}	-0.349	0.001	0.001	β_{31}	-0.372	-0.022	0.003
700	β_{02}	0.227	-0.123	0.018	β_{02}	0.122	-0.228	0.053
	β_{03}	0.812	-0.038	0.003	β_{03}	0.513	-0.337	0.115
	β_{04}	-0.948	0.002	0.025	β_{04}	-0.968	-0.018	0.026
	β_{14}	0.302	0.002	0.044	β_{14}	0.295	-0.005	0.046

Table 5.3. The AEs, biases and MSEs for the parametric and semiparametric GOLLMaxM regression models based on 1,000 simulations for scenario 2.

	Semiparametric				Parametric			
n	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.197	-0.003	0.005	β_{21}	0.218	0.018	0.010
	β_{31}	-0.316	0.034	0.011	β_{31}	-0.297	0.053	0.018
350	β_{02}	0.232	-0.118	0.024	β_{02}	0.129	-0.221	0.054
	β_{03}	0.883	0.033	0.009	β_{03}	0.579	-0.271	0.080
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.204	0.004	0.001	β_{21}	0.211	0.011	0.002
	β_{31}	-0.351	-0.001	0.003	β_{31}	-0.309	0.041	0.007
350	β_{20}	0.433	-0.017	0.021	β_{02}	0.335	-0.225	0.068
	β_{03}	1.293	0.093	0.020	β_{03}	0.848	-0.352	0.139
	Semiparametric				Parametric			
	Parameter	AE	Bias	MSE	Parameter	AE	Bias	MSE
	β_{21}	0.200	0.000	0.001	β_{21}	0.201	0.001	0.001
	β_{31}	-0.342	0.008	0.003	β_{31}	-0.351	-0.001	0.005
700	β_{02}	0.225	-0.125	0.020	β_{02}	0.111	-0.239	0.059
	β_{03}	0.819	-0.031	0.003	β_{03}	0.552	-0.328	0.110

on the simulation results in Tables 5.2 and 5.3, we interest is in verifying how much the inclusion of an additive term affects in the estimations of the other fixed parameters. For semiparametric model, we verify that the MSEs of the MLEs of β_{21}, β_{31}, β_{02}, β_{03}, β_{04} and β_{14} for scenario 1 and 2 decay toward zero when the sample size n increases, as usually expected under first-order asymptotic theory. The mean estimates of the parameters tend to be closer to the true parameter values when n increases. However, for the parametric model, such measures do not exhibit the same behavior.

In relation to the behavior of the nonlinear effects in the simulations (scenarios 1 and 2), in

the Figures 5.5 and 5.6 displays the generated and fitted effects for the parametric and semi-parametric models. We also present in this figure the box-plots of the GD, AIC and BIC statistics obtained in 1,000 simulations for both models. We can note that the nonlinear effects are very close to the true shape as shown in the Figure 5.4, when the sample size increases. Further, we can conclude that the semi-parametric model presents the lowest values of GD, AIC and BIC statistics, indicating that it is the most suitable model for the simulated data.

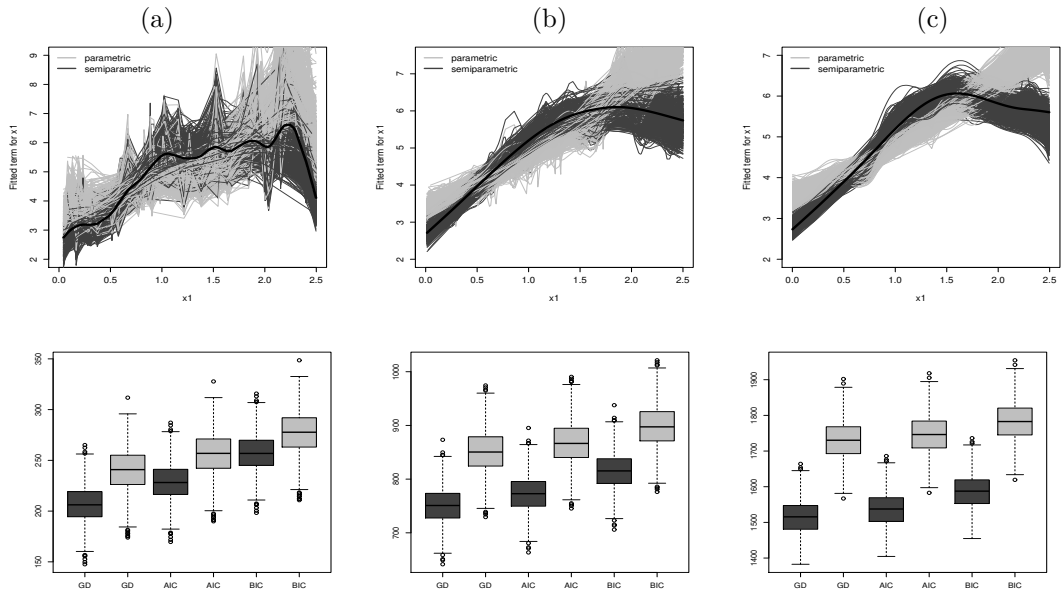

Figure 5.5. The fitted GOLLMaxM semiparametric and parametric regression model under scenario 1 of the effect of x_1 in μ whit goodness-of-fit statistics measures. (a) For $n = 100$. (b) For $n = 350$. (c) For $n = 700$.

5.5 Applications

In this section, we apply the GOLLMaxM model to prostate cancer data. The patient data come from a study developed by Kattan *et al.* (1999) and Stephenson *et al.* (2005) at the Cleveland Clinic. The study cohort comprises 1324 patients with clinically localized prostate cancer treated by open radical prostatectomy between 1987 and 2003. Patient data were obtained from the Cleveland Clinic from a single surgeon and the authors have the rights on the data. Patients with clinical stage T1a or T1b disease, who received neoadjuvant therapy or adjuvant therapy or who had missing data for PSA were excluded. All information was obtained with appropriate Institutional Review Board waivers. The response variable is given by the number of months (t_i) without detectable disease after prostatectomy. Uncensored observations correspond to patients having the cancer recurrence time computed. Censoring observations correspond to patients who were not observed to have cancer recurrence at the time the data were collected. The numbers of censoring and uncensored observations are 1096 and 228, respectively, of the total of 1324 patients. The following explanatory variables were associated with each patient (for $i = 1, \ldots, 1324$):

- δ_i is the event indicator where 1 represents the event and 0 is censored;
- x_{i1} is the PSA value (in ng/ml) from the laboratory report before undergoing prostatectomy;

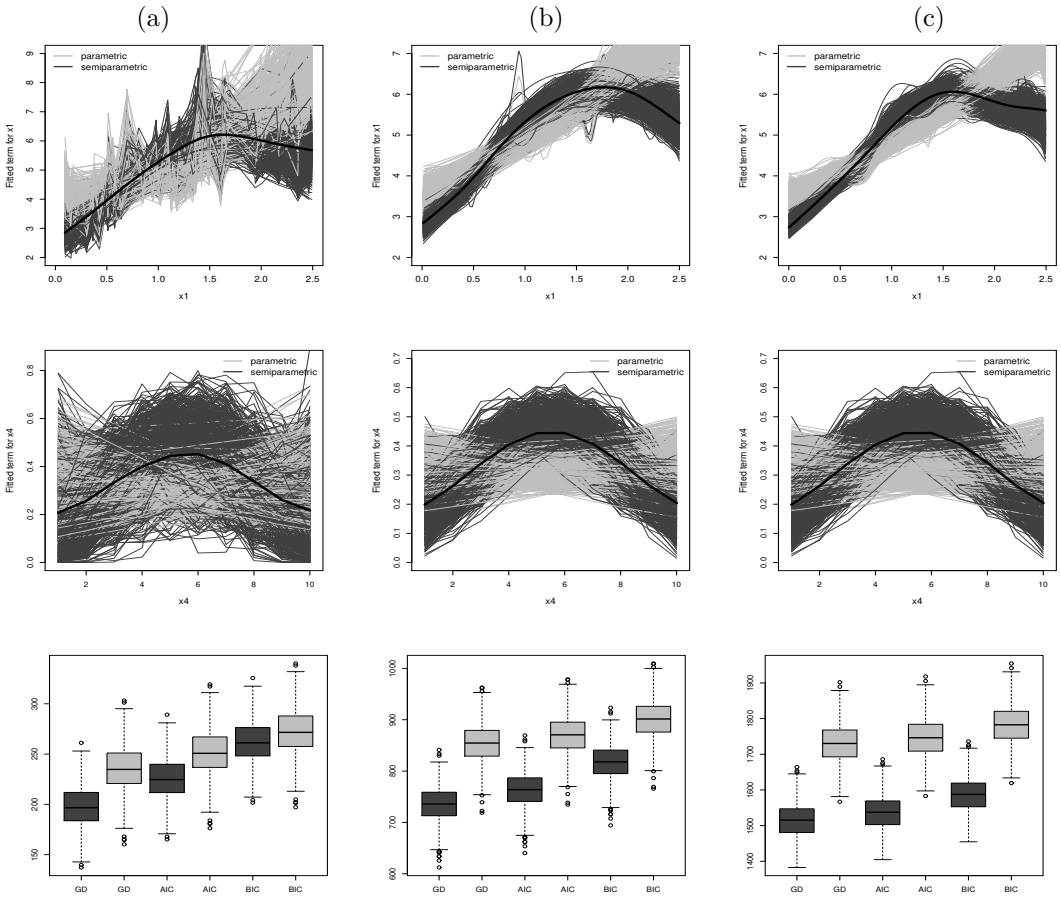

Figure 5.6. The fitted GOLLMaxM semiparametric and parametric regression model under scenario 2 of the effect of x_1 in μ and x_4 in τ whit goodness-of-fit statistics measures. (a) For $n = 100$. (b) For $n = 350$. (c) For $n = 700$.

- x_{i2} is whether the patient received neoadjuvant hormones, that is, treated with hormone therapy prior to radical prostatectomy ($x_{i2} = 1$, yes and $x_{i2} = 0$, no);
- x_{i3} is the extracapsular extension on path report ($x_{i3} = 1$, yes and $x_{i3} = 0$, no);
- x_{i4} is the seminal vesicle invasion on path report ($x_{i4} = 1$, yes and $x_{i4} = 0$, no);
- x_{i5} is the lymph node involvement on path report ($x_{i5} = 1$, negative and $x_{i5} = 0$, positive);
- x_{i6} is surgical margin status ($x_{i6} = 1$, yes and $x_{i6} = 0$, no).

5.5.1 Descriptive analysis of the prostat data

Next, we perform an exploratory analysis of these data. The percentage of individuals considered cured is approximately 83%. In Figure 5.7 we present the empirical survival function by explanatory variables. In particular attention to the variable lymph node involvement on path report, x_{i5}, for level $x_{i5} = 0$, there are 7 patients who did not have a recurrence of the disease and 26 recurring; for level $x_{i5} = 1$, 1089 did not have a recurrence and 202 had a recurrence of prostate cancer.

We considered a marginal analysis considering only the time response variable. In this study we consider the proposed model GOLLMax2 and Weibull mix model. the values of GD, AIC and BIC

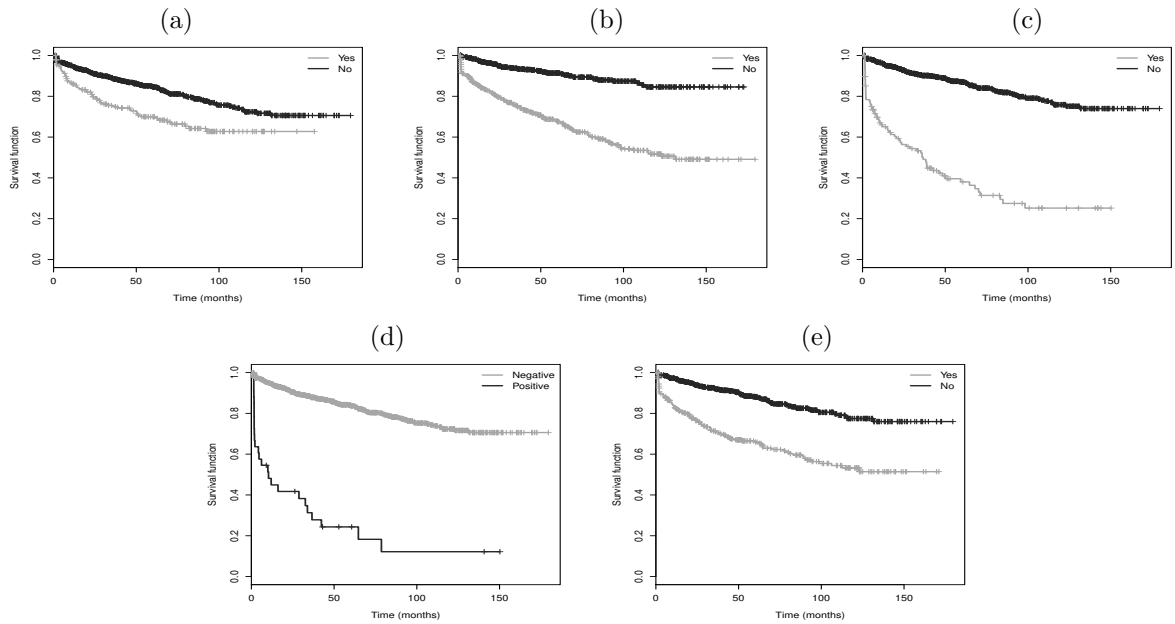

Figure 5.7. Plots of the empirical survival curves for prostate cancer data.(a) By x_{i2}. (b) By x_{i3}. (c) By x_{i4}. (d) By x_{i5}. (e) By x_{i6}.

statistics are given in Table 5.4. We note that the GOLLMaxM model presents the lowest values of statistics, that is, the GOLLMaxM model adjusts better the prostate cancer data compared to the Weibull mixture model.

Table 5.4. The GD, AIC and BIC measurements for GOLLMaxM and Weibull mixture models for prostate cancer data.

Models	GD	AIC	BIC
GOLLMaxM	2998.5	3006.5	3027.3
Weibull mixture	3006.2	3012.2	3027.7

In order to evaluate if the model is appropriate, Figure 5.8a and 5.8b adjusted curves of the survival and hazard functions, respectively, considered under the marginal analysis of the response variable t_i.

5.5.2 Semiparametric regression model

Next, we present results from the fit of the regression (5.6) with two systematic components:

$$\mu_i = \exp\left(\beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2} + \beta_{31}x_{i3} + \beta_{41}x_{i4} + \beta_{51}x_{i5} + \beta_{61}x_{i6}\right)$$

and

$$\tau_i = \frac{\exp\left(\beta_{04} + cs(x_{i1}, df^* = 3) + \beta_{24}x_{i2} + \beta_{34}x_{i3} + \beta_{44}x_{i4} + \beta_{54}x_{i5} + \beta_{64}x_{i6}\right)}{1 + \exp\left(\beta_{04} + cs(x_{i1}, df^* = 3) + \beta_{24}x_{i2} + \beta_{34}x_{i3} + \beta_{44}x_{i4} + \beta_{54}x_{i5} + \beta_{64}x_{i6}\right)}.$$

The MLEs, SEs and p-value associated for the GOLLMaxM and Weibull mixture models are presented in Table 5.5. Thus, when establishing a significance level of 5%, we note that the voltage level is significant and should be used to model the location and scale.

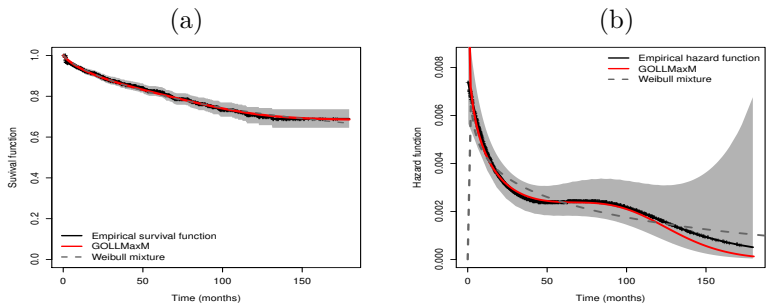

Figure 5.8. Plots for prostate cancer data. (a) Empirical survival function and estimated GOLLMaxM and Weibull mixture distributions. (b) Empirical hazard function and estimated GOLLMaxM and Weibull mixture distributions.

For the μ_i regression structure, the variable x_{i3} is not significant for the GOLLMaxM model and in the Weibull mixture model all variables are significant. In the regression structure of the proportion of cured τ_i it appears that x_{i2} and x_{i5} are not significant for both models. The coefficient of the nonlinear term is presented, however such value is not interpreted.

Table 5.5. MLEs, SEs and p-values for the LOLLGG regression model fitted to the voltage level data.

GOLLMaxM				Weibull mixture			
Parameter	Estimate	SE	p-Value	Parameter	Estimate	SE	p-Value
$\log(\sigma)$	-0.494	0.033	-	$\log(\sigma)$	-0.145	0.020	-
$\log(\nu)$	-0.690	0.017	-	-	-	-	-
β_{01}	4.404	0.040	< 0.001	β_{01}	4.870	0.241	< 0.001
β_{11}	-0.008	0.001	< 0.001	β_{11}	-0.020	0.004	< 0.001
β_{21}	-0.754	0.026	< 0.001	β_{21}	-0.242	0.091	0.008
β_{31}	0.019	0.030	0.524	β_{31}	0.432	0.073	< 0.001
β_{41}	-0.448	0.034	< 0.001	β_{41}	-1.696	0.128	< 0.001
β_{51}	0.166	0.019	< 0.001	β_{51}	0.574	0.228	0.012
β_{61}	-0.644	0.027	< 0.001	β_{61}	-0.725	0.076	< 0.001
β_{04}	1.127	1.146	0.325	β_{03}	1.590	0.966	0.100
$cs(x1)$	-0.059	0.008	< 0.001	$cs(x1)$	-0.054	0.009	< 0.001
β_{24}	-0.176	0.114	0.122	β_{23}	-1.018	0.165	< 0.001
β_{34}	-1.106	0.097	< 0.001	β_{33}	-2.429	0.209	< 0.001
β_{44}	-1.187	0.314	0.0001	β_{43}	0.548	0.299	0.0678
β_{54}	1.137	1.143	0.319	β_{53}	0.391	0.960	0.684
β_{54}	-0.603	0.102	< 0.001	β_{63}	-0.751	0.130	< 0.001
GD=2729.16	AIC=2767.16	BIC=2865.73		GD=2729.59	AIC=2773.59	BIC=2887.72	

In Figure 5.9, we perform the residual analysis by plotting the quantile residuals $\hat{rq}_i$ (see Subsection 5.3.3) against the index of observations for the fitted GOLLMaxM semiparametric regression models in the interval $[-3, 3]$.

Figures 5.10 and 5.11 give the plots of the quantile residuals by density, qq-plot and worm plot for the GOLLMaxM and Weibull mixture models, respectively, to detect possible outlying observations as well as departures from the assumptions of semiparametric regression models. Note that Figures 5.11b and c show normality deviations and extrapolation with respect to the confidence bands in the worm plots.

The partial effects for the covariates in relation to the systematic structures are displayed in Figures 5.12. In Figure 5.12a, we present the effect of the term linear for variable x_1, as we have already interpreted before, when the age of the patient is elevated at the moment of diagnosis the lifetime

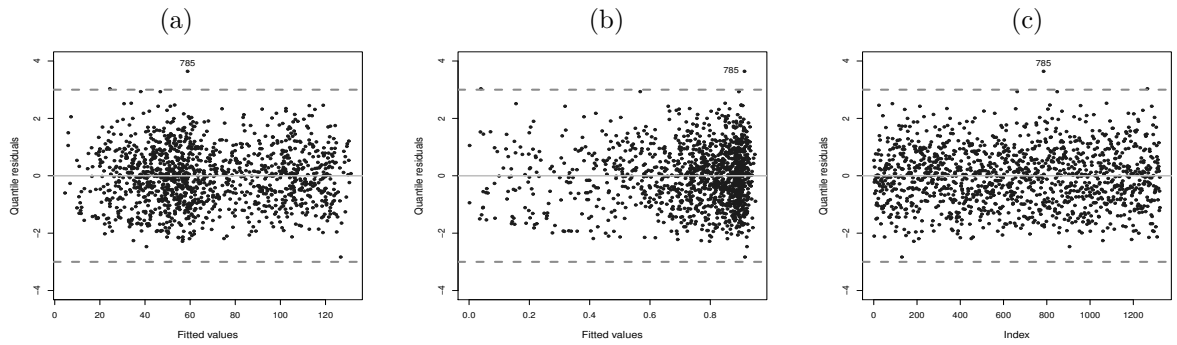

Figure 5.9. Plots for GOLLMaxM semiparametric regression model. (a) Fitted (whit regression structure in μ_i) versus quantile residuals. (b) Fitted (whit regression structure in τ_i) versus quantile residuals. (c) Index plot of the quantile residual for the prostate cancer data.

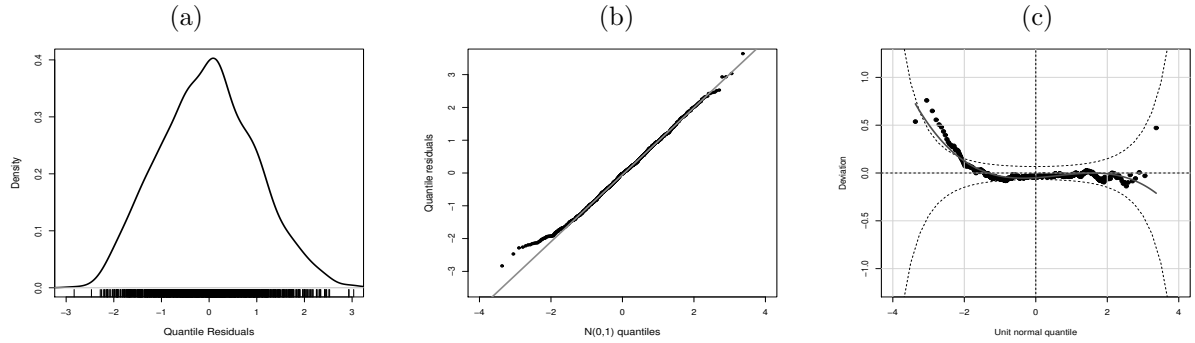

Figure 5.10. Plots of the GOLLMaxM semiparametric regression model for prostate cancer data. (a) Density of the quantile residuals. (b) Q-Q plot for quantile residuals. (c) Worm plot for quantile residual.

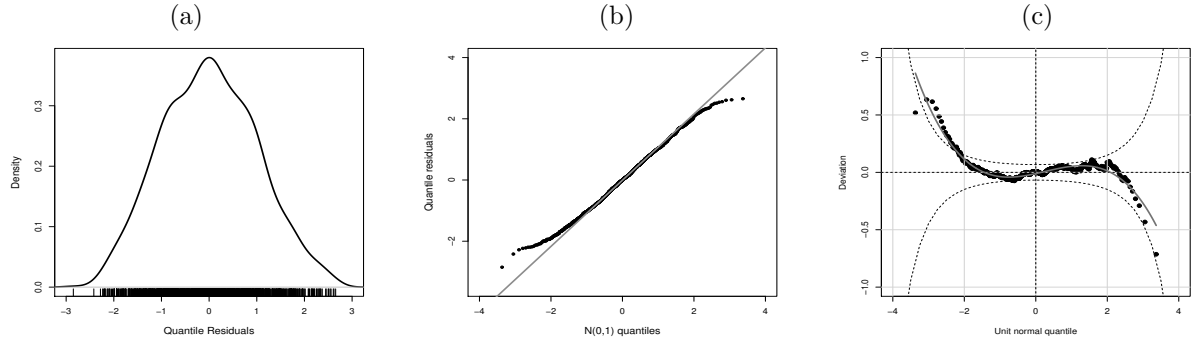

Figure 5.11. Plots of the Weibull mixture semiparametric regression model for prostate cancer data. (a) Density of the quantile residuals. (b) Q-Q plot for quantile residuals. (c) Worm plot for quantile residual.

decreases. Figure 5.12b indicates that for monoclonal protein spike measurements between approximately 0.5 and 1 there is a growing linear trend in lifetime. For values between 1 and 2.6 the lifetime it remains constant and from 2.6 there is a linear decreasing trend in the lifetime.

Equation (5.6) provides an estimate of the cure probability for any prostate cancer patient who underwent radical prostatectomy in terms of the above explanatory variables:

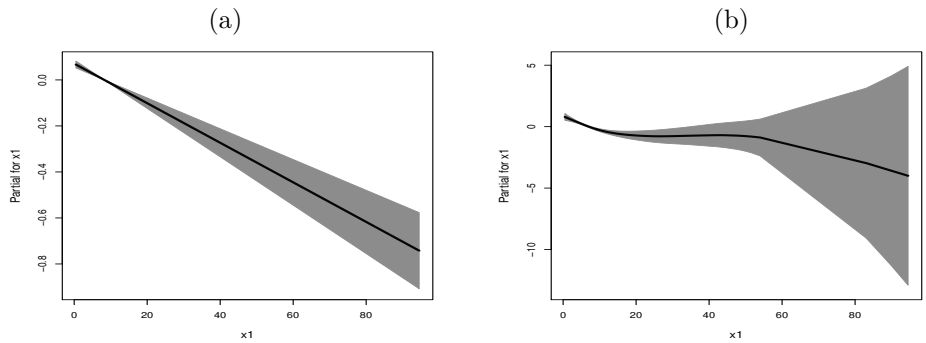

Figure 5.12. The GOLLMaxM semiparametric regression model fitted for the prostate cancer data. (a) Fitted partial effects of the x_{i1} for structure regression in μ_i. (b) Fitted partial effects of the x_{i1} for structure regression in τ_i.

$$\hat{\tau}_i = \frac{\exp\left(\hat{\beta_{04}} + cs(x_{i1}, df^* = 3) + \hat{\beta_{24}}x_{i2} + \hat{\beta_{34}}x_{i3} + \hat{\beta_{44}}x_{i4} + \hat{\beta_{54}}x_{i5} + \hat{\beta_{64}}x_{i6}\right)}{1 + \exp\left(\hat{\beta_{04}} + cs(x_{i1}, df^* = 3) + \hat{\beta_{24}}x_{i2} + \hat{\beta_{34}}x_{i3} + \hat{\beta_{44}}x_{i4} + \hat{\beta_{54}}x_{i5} + \hat{\beta_{64}}x_{i6}\right)}. \tag{5.11}$$

Then, the overall cure probability for the population under study is obtained from Equation (5.11)

$$\hat{\tau} = \frac{\sum_{i=1}^{1324} \hat{\tau}_i}{1324} = 0.71.$$

Hence, for the total population of the 1324 patients with localized prostate cancer treated by open radical prostatectomy, the estimate of the cure fraction is about $\hat{\tau} = 0.71$.

Considering the following scenario $x_{i1} = 15$, $x_{i2} = 1$, $x_{i3} = 1$, $x_{i4} = 1$, $x_{i5} = 1$ and stratification by variable $x_{i6} = 1$, is presented the plots comparing the empirical and estimated survival functions for the GOLLMaxM semiparametric regression model are given in Figure 5.13a. In Figure 5.13b, we also present the fitted hazard functions. From these plots, we can note a significant different between the survival curves by scenario considered. These plots indicate that the GOLLMaxM regression model yields a satisfactory fit to the current data.

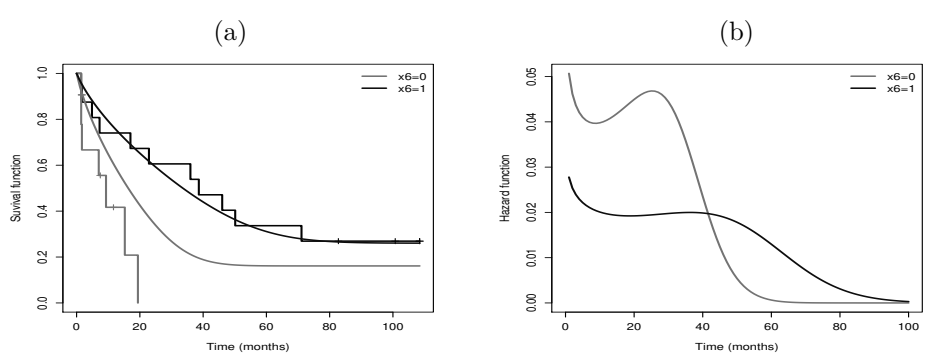

Figure 5.13. Plots for fixed values $x_{i1} = 15$, $x_{i2} = 1$, $x_{i3} = 1$, $x_{i4} = 1$, $x_{i5} = 1$ and stratification by variable $x_{i6} = 1$.(a) Empirical and estimated survival functions for the GOLLMaxM semiparametric regression model. (b) Empirical and estimated hazard functions for the GOLLMaxM semiparametric regression model.

5.6 Concluding Remarks

We define a new *generalized odd log-logistic Maxwell mixture* (GOLLMaxM) semiparametric regression to analyze data in presence of cure rate, heteroskedasticity and nonlinear effects in covariables. We discuss some inferential issues related to this regression and perform some simulations. We illustrate the potentiality of the new regression by means of prostate cancer data which includes descriptive and marginal analysis, model checking and interpretations of its systematic components. In conclusion, the proposed semiparametric regression is a effective alternative in demonstrating all the biological and productive effects observed in the field. In conclusion, the proposed semiparametric regression is an alternative in data modeling in the presence of cure rate.

References

Berkson, J.; Gage, R.P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**, 501-515.

Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, textbf11, 15-53.

Cox, D.R. Regression models and life tables (with discussion). **Journal of the Royal Statistical Society**, v. 34, n. 2, p. 187-220, 1972.

Kattan, M.W., Wheeler, T.M. and Scardino, P.T. (1999). Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *Journal of clinical oncology*, **17**, 1499-1499.

Li C.S, Taylor J.M and Sy J.P. (2001). Identifiability of cure models. *Statistics & Probability Letters*, **54**, 389?395.

Martinez, E.Z., Achcar, J.A., Jácome, A.A. and Santos, J.S. (2013). Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data. *Computer methods and programs in biomedicine*, **112**, 343-355.

Maller R. and Zhou X. (1996). *Survival Analysis with Long-term Survivors*, Wiley, New York, NY.

Ortega, E.M.M., Cordeiro, G.M., Kattan, M.W. (2013). The log-beta Weibull regression model with application to predict recurrence of prostate cancer. *Statistical Papers*, **54**, 113-132.

Ortega, E.M.M., Cordeiro, G.M., Kattan, M.W. (2012). The negative binomial?beta Weibull regression model to predict the cure of prostate cancer. *Journal of Applied Statistics*, **39**, n. 6, p. 1191-1210, 2012.

Ortega, E.M.M., Cancho, V.G., Paula, G.A. (2009). Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis*, **15**, p. 79.

Ramires, T. G., Hens, N., Cordeiro, G. M. and Ortega, E. M. (2018). Estimating nonlinear effects in the presence of cure fraction using a semi-parametric regression model. *Computational Statistics*, **33**, 709-730.

Rodrigues, J., Cancho, V.G., de Castro, M. and Louzada-Neto, F. (2009). On the unification of long-term survival models. *Statistics & Probability Letters*, **79**, 753-759.

Stephenson, A.J., Scardino, P.T., Eastham, J.A., Bianco Jr, F.J., Dotan, Z.A., DiBlasio, C.J., Reuther, A., Klein, E.A. and Kattan, M.W. (2005). Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, **23**, p. 7005.

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

Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., Voudouris V. and De Bastiani, F. (2017), *Flexible Regression and Smoothing: Using GAMLSS in R*, Chapman and Hall/CRC, New York.

Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J. and Stasinopoulos, D. (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, **39**, 1279-1293.

6 THE GENERALIZED ODD LOG-LOGISTIC MAXWELL REPARAMETRIZED

We propose a reparametrization regression model considering the GOLLMax distribution. The proposed model features a parameter related to the median. The reparametrized distribution can be used in different contexts, for example, in experimental data when the response variable belongs to the positive reals. Thus, a median regression model is proposed. Parameter estimates are obtained by maximum likelihood. Some global-influence measurements and quantile residuals are also investigated. The proposals are illustrated by one application in experimental data.
Keywords: Experimental Data; Diagnostics; Likelihood inference ; Median; Regression model.

6.1 Introduction

Defining new families by adding shape parameters to control skewness, kurtosis and tail weights, providing great flexibility in modeling skewed data in practice, is an important focus in statistics in several papers as, for example, Mudholkar and Srivastava (1993) present the exponentiated Weibull distribution and applications, Cordeiro, Ortega, and Silva (2011) defined the exponentiated generalized gamma (EGG) distribution. da Cruz et al. (2016) proposed the odd log-logistic Weibull distribution, Braga *et al.* (2016) studied the odd log-logistic normal distribution. Recently, Prataviera *et al.* (2018) defined the heteroscedastic odd log-logistic generalized gamma regression model for censored data and the generalized odd log-logistic family studied by Cordeiro *et al.* (2017), among others.

The great difficulty when generalizing a distribution is the interpretation of the parameters when a regression structure is considered. For example, in Braga, the base distribution is the normal one in which the parameters represent the mean and variance, respectively. However, in the model obtained by the generalization, the parameters no longer have the original characteristic. Thus, this work proposes a reparametrization of the GOLLMax in terms of the median.

This chapter is organized as follows. In Section 6.2, we define the reparametrized GOLLMax. In Section 6.3, we define the median regression model. In Section 6.4, we provide one application to real data to illustrate the flexibility of the model. Section 6.5 offers some concluding remarks and future works.

6.2 The model definition

A random variable Y has the GOLLMax distribution if its cumulative distribution function (cdf) and probability density function (pdf) are (for $y > 0$)

$$F(y;\alpha,\sigma,\nu) = \frac{\gamma_1^{\nu\,\sigma}(3/2, y^2/\alpha^2)}{\gamma_1^{\nu\sigma}(3/2, y^2/\alpha^2) + [1-\gamma_1^{\sigma}(3/2, y^2/\alpha^2)]^{\nu}} \tag{6.1}$$

and

$$f(y;\alpha,\sigma,\nu) = \frac{4\,\nu\,\sigma\,y^2}{\sqrt{\pi}\,\alpha^3}\,\exp\left(-\frac{y^2}{\alpha^2}\right)\frac{\gamma_1^{\nu\sigma-1}(3/2, y^2/\alpha^2)\left[1-\gamma_1^{\sigma}(3/2, y^2/\alpha^2)\right]^{\nu-1}}{\{\gamma_1^{\nu\sigma}(3/2, y^2/\alpha^2) + [1-\gamma_1^{\sigma}(3/2, y^2/\alpha^2)]^{\nu}\}^2}, \tag{6.2}$$

respectively, where $\nu > 0$ and $\sigma > 0$ are two extra shape parameters and $\alpha > 0$ is scale parameter. The $\gamma_1(p,u) = \gamma(p,u)/\Gamma(p)$ is the incomplete gamma function ratio, $\gamma(p,u) = \int_0^u w^{p-1}\mathrm{e}^{-\mathrm{w}}\mathrm{dw}$ is the incomplete gamma function and $\Gamma(\cdot)$ is the gamma function. Henceforth, if Y is a random variable with cdf (6.1), we write $Y \sim \mathrm{GOLLMax}(\alpha,\nu,\sigma)$.

The GOLLMax distribution can be simulated by inverting (6.1). In fact, the quantile function (qf) of Y can be expressed as

$$y = Q_{\text{Max}}\,(w;\alpha)\,,$$

where $w = \left[\frac{u^{\frac{1}{\nu}}}{(1-u)^{\frac{1}{\nu}}+u^{\frac{1}{\nu}}}\right]^{\frac{1}{\sigma}}$ and $Q_{\text{Max}}(w;\alpha) = G^{-1}(w;\alpha)$ is the qf of the Maxwell distribution.

Further, we can write

$$y = Q_{\text{Max}}(w;\alpha) = \alpha\,\sqrt{\gamma_1^{-1}(3/2,w)}, \tag{6.3}$$

where $\gamma_1^{-1}(3/2,w)$ is the inverse of the upper gamma regularized function. For more details, see `http://functions.wolfram.com/GammaBetaErf/InverseGammaRegularized/`.

The median, first and third quartiles of GOLLMax distribution are

$$\mu = Q_{0.50} = \alpha\,\sqrt{\gamma_1^{-1}(3/2,0.5^{\frac{1}{\sigma}})},$$

$$Q_{0.25} = \alpha\,\sqrt{\gamma_1^{-1}\left(3/2,\left[1/(3^{\frac{1}{\nu}}+1)\right]^{\frac{1}{\sigma}}\right)} \quad \text{end} \quad Q_{0.75} = \alpha\,\sqrt{\gamma_1^{-1}\left(3/2,\left[3^{\frac{1}{\nu}}/(3^{\frac{1}{\nu}}+1)\right]^{\frac{1}{\sigma}}\right)}.$$

The proposed reparametrization of GOLLMax models is based on the median, $Q_{0.50}$. The relation of the new parameter μ is given as follows

$$\alpha = \frac{\mu}{\sqrt{\gamma_1^{-1}(3/2,0.5^{\frac{1}{\sigma}})}},$$

where $\mu > 0$ is scale and position parameter. A random variable Y has the GOLLMax reparametrized distribution if its pdf is

$$f(y;\mu,\sigma,\nu) = \frac{4\,\nu\,\sigma\,b_\sigma^{3/2}y^2}{\sqrt{\pi}\,\mu^3}\,\exp\left(-\frac{b_\sigma\,y^2}{\mu^2}\right)\frac{\gamma_1^{\nu\sigma-1}(3/2,b_\sigma\,y^2/\mu^2)\left[1-\gamma_1^{\sigma}(3/2,b_\sigma\,y^2/\mu^2)\right]^{\nu-1}}{\left\{\gamma_1^{\nu\sigma}(3/2,b_\sigma\,y^2/\mu^2)+\left[1-\gamma_1^{\sigma}(3/2,b_\sigma\,y^2/\mu^2)\right]^{\nu}\right\}^2}. \tag{6.4}$$

where $b_\sigma = \gamma_1^{-1}(3/2,0.5^{\frac{1}{\sigma}})$. If Y is a random variable with cdf (6.4), we denote $Y \sim \text{GOLLMax2}(\mu,\sigma,\nu)$.

Some possible shapes of the density (6.4) for selected parameter values, including well-known distributions, are displayed in Figure 6.1, thus emphasizing its asymmetrical, symmetrical and lightly bimodal shapes. Note that the additional shape parameters give flexibility to the new distribution. We have lightly bimodality when $\nu \in (0,1)$, among other shapes depending on the parameter values.

The GOLLMax2 distribution can be simulated by replacing α in (6.3). The qf of Y can be expressed as

$$y = Q_{\text{Max}}(w;\mu,\sigma,\nu) = \mu\,\frac{\sqrt{\gamma_1^{-1}(3/2,w)}}{\sqrt{\gamma_1^{-1}(3/2,0.5^{\frac{1}{\sigma}})}}, \tag{6.5}$$

6.3 GOLLMax2 median regression model

In this section, we define a reparametrized GOLLMax regression model based in median as presented in the Section 6.2. Regression analysis has a potential interest in verifying and estimating the effects of one or more covariables in relation to a given distribution parameters to be used. Such relation

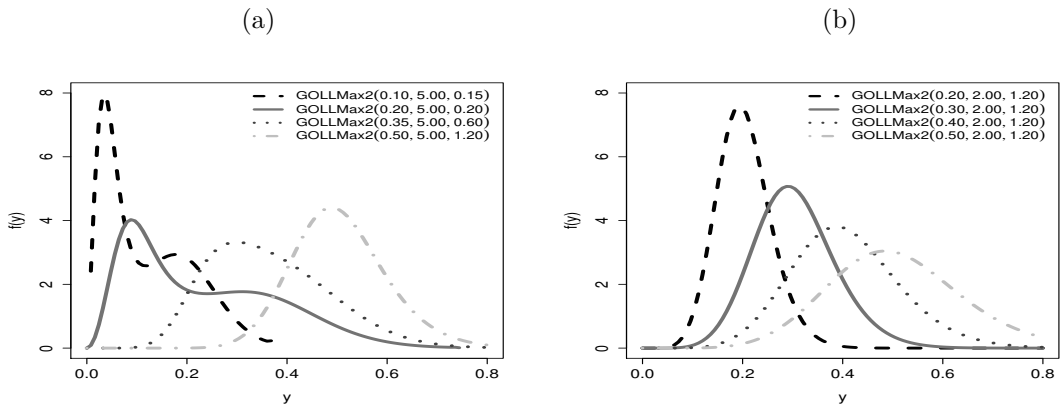

Figure 6.1. Plots of the GOLLMax2 density for selected parameter values.(a) For asymmetrical and lightly bimodal shapes. (b) For symmetrical shapes.

can be in relation to the average, median or some quantiles. In this work, we are interested in studying the median regression model. The systematic components for μ parameter in density (6.4) to allow them varying across the observations (for $i = 1, \ldots, n$) as

$$g(\mu_i) = \mathbf{v}_i^T \boldsymbol{\beta}, \tag{6.6}$$

where $g : [0, \infty) \to R$ is known one-to-one link functions continuously twice differentiables, $\mathbf{v}_i^T = (v_{i1}, \ldots, v_{ip})$ is a vector of known explanatory variables for the ith observation, and $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{1p})^T$ is a parameter vectors of dimension p. Then, $g(\boldsymbol{\mu}) = \mathbf{V}\boldsymbol{\beta}$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)^T$ is a specified $n \times p$ matrix of full column rank $p < n$. The usual systematic component for the scale parameter is $g(\mu_i) = \log(\mu_i)$, then μ_i can be obtained by inverting $g(\mu_i)$, as $\mu_i = \exp(\mathbf{v}_i^T \boldsymbol{\beta})$.

Consider a sample $(y_1, \mathbf{v}_1), \ldots, (y_n, \mathbf{v}_n)$ of n independent observations. Conventional likelihood estimation techniques can be applied here. The total log-likelihood function for the vector of parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \sigma, \nu)^T$ from model (6.4) takes the form

$$\begin{aligned} l(\boldsymbol{\psi}) \;=\; & n \log\left(\frac{4}{\sqrt{\pi}}\right) + \frac{3}{2}\log(b_\sigma) + \sum_{i=1}^{n}\log(\nu) + \sum_{i=1}^{n}\log(\sigma) + \sum_{i=1}^{n}\log\left(\frac{b_\sigma\, y_i^2}{\mu_i^3}\right) - \sum_{i=1}^{n}\left(\frac{b_\sigma\, y_i^2}{\mu_i^2}\right) + \\ & (\sigma\,\nu - 1)\sum_{i=1}^{n}\log\left[\gamma_1(3/2, b_\sigma\, y_i^2/\mu_i^2)\right] + (\nu - 1)\sum_{i=1}^{n}\log\left[1 - \gamma_1(3/2, b_\sigma\, y_i^2/\mu_i^2)\right] - \\ & 2\sum_{i=1}^{n}\log\left\{\gamma_1^{\sigma_i\,\nu}(3/2, b_\sigma\, y_i^2/\mu_i^2) + \left[1 - \gamma_1^{\sigma}(3/2, b_\sigma\, y_i^2/\mu_i^2)\right]^{\nu}\right\}. \end{aligned} \tag{6.7}$$

The MLE $\widehat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ can be calculated by maximizing the log-likelihood (6.7). The numerical maximization of (6.7) can be done in the `gamlss` packages in `R` using the `RS` maximization algorithm (Rigby and Stasinopoulos, 2007) to determine the estimate $\hat{\boldsymbol{\psi}}$.

For selection of the appropriate distribution, we use the global deviance (GD), say GD $= -2\mathrm{l}(\widehat{\boldsymbol{\theta}})$, where $l(\hat{\boldsymbol{\theta}})$ is the maximized log-likelihood function (6.7), and the generalized Akaike information criterion (GAIC) given by GAIC(k) = GD + k $\times$ df, where *df* is the total degrees of freedom of the adjusted model and k is the penalty for each degree of freedom used. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are special cases of the GAIC(k) measure when $k = 2$ and $k = \log(n)$, respectively. We consider the (GD), AIC and BIC criteria to select the best regressions.

6.3.1 Influence and residual analysis

We adopt diagnostic measures based on case deletion (global influence) to detect influential observations in the proposed regression model. The case-deletion model with regression model structure (6.6). Thus, the log-likelihood function for $\boldsymbol{\psi}$ is denoted by $l_{(i)}(\boldsymbol{\psi})$. A global influence measure considered by (Xie and Wei, 2007) is a generalization of the Cook distance defined as a standardized norm of $\hat{\boldsymbol{\psi}}_{(i)} - \hat{\boldsymbol{\psi}}$ (generalized Cook distance), namely

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

The measure to evaluate the influence of an observation (Cook and Weisberg, 1982), called the log-likelihood distance, is the difference between $\hat{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\omega}}_{(i)}$ on the log-likelihood scale, namely

$$LD_i = 2\Big[l(\hat{\boldsymbol{\psi}}) - l(\hat{\boldsymbol{\psi}}_{(i)})\Big], \tag{6.8}$$

where $l(\hat{\boldsymbol{\psi}})$ is the maximized log-likelihood for the full sample and $l(\hat{\boldsymbol{\psi}}_{(i)})$ is the maximized log-likelihood for the sample excluding the ith observation.

Another important step in the analysis of a fitted model is to check possible deviations from the model assumptions. In this context, we consider the quantile residuals (Dunn and Smyth, 1996) for the GOLLMax2 median regression model, is defined by

$$\hat{qr}_i = \Phi^{-1}\left\{\frac{\gamma_1^{\hat{\sigma}\hat{\nu}}(3/2, \hat{b_\sigma}\, y_i^2/\hat{\mu}_i^2)}{\gamma_1^{\hat{\sigma}\hat{\nu}}(3/2, \hat{b_\sigma}\, y_i^2/\hat{\mu}_i^2) + \left[1 - \gamma_1^{\hat{\sigma}}(3/2, \hat{b_\sigma}\, y_i^2/\hat{\mu}_i^2)\right]^{\hat{\nu}}}\right\}, \tag{6.9}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative standard normal distribution.

We built envelopes to enable better interpretation of the probability normal plot of the residuals. These envelopes are simulated confidence bands described by Atkinson (1985) that contain the residuals, such that if the model is well-fitted, the majority of points will be randomly distributed within these bands.

6.4 Applications

In this section, we provide two applications to real data sets to prove empirically the flexibility of the proposed models. The calculations are performed using the `gamlss` packages in the `R` software. To compare the performance of the proposed regression model, we consider some distributions. We consider the gamma (GA) distribution with density function given by

$$f(y, \mu, \sigma) = \frac{y^{\left(\frac{1}{\sigma^2}-1\right)}}{(\sigma^2\mu)^{\frac{1}{\sigma^2}}\,\Gamma\left(\frac{1}{\sigma^2}\right)} \exp\left\{-\frac{y}{\sigma^2\mu}\right\}, \quad y > 0, \tag{6.10}$$

where $\mu > 0$ and $\sigma > 0$. In this parameterization we have, $E[Y] = \mu$ and variance $Var[Y] = \mu^2\sigma^2$, such that σ is the square root of the dispersion parameter of the usual GLM gamma model.

The three parameter generalized inverse Gaussian (GIG) defined by Jørgensen (1982) has density function ($y > 0$) given by

$$f(y, \mu, \sigma, \nu) = \left(\frac{b}{\mu}\right)^{\nu}\left[\frac{y^{\nu-1}}{2k_\nu\left(\frac{1}{\sigma^2}\right)}\right]\exp\left\{-\frac{1}{2\sigma^2}\left(\frac{b\,y}{\mu} + \frac{\mu}{b\,y}\right)\right\}, \quad y > 0, \tag{6.11}$$

where $\mu > 0$, $\sigma > 0$, $\nu \in \Re$, $b = \left[k_{\nu+1}\left(\frac{1}{\sigma^2}\right)\right]\left[k_\nu\left(\frac{1}{\sigma^2}\right)\right]^{-1}$ and $k_\lambda(t) = \frac{1}{2}\int_0^\infty y^\lambda \exp\left\{-\frac{1}{2}t(y + y^{-1})\right\} dy$. The μ is the mean and scaling parameter and the variance is $Var[Y] = \mu^2\left[\frac{2\sigma^2}{b}(\nu+1) + \frac{1}{b^2} - 1\right]$.

We also consider the normal (NO) distribution. However, for comparison purposes the link function logarithmic is used for the parameters μ.

6.4.1 Application 1: Effect of doses

The first data set comes from an experiment carried out to assess the effects of doses of an anthelmintic compound (ml) to control a parasite (fixed effects) using a CRD with five treatments. Treatment 1 and treatment 2 are controls, and treatment 3, treatment 4, and treatment 5 use a new drug at concentrations of 5%, 10%, and 15%, with six replications. The data set with $n = 30$ observations is available at Professor Euclides Malheiros Braga's website: `http://jaguar.fcav.unesp.br/euclides/`. Choose the year 2013 and the option Estatistica Experimental-PG em Ciencia Animal (UFERSA) and download the file A DIC ex2.txt. These data were recently described and analyzed by Braga *et al.* (2016). The authors propose a new distribution based on the extension of the normal distribution.

Table 6.1 summarizes the main descriptive statistics for each of the five treatments and suggest positively skewed except for Treatment 1 distributions with different degrees of variability, skewness and kurtosis. We can also verify by Figure 6.2, that treatments 1 and 2 are quite distant in relation to treatments 3, 4 and 5 considering the measures of the mean and median.

Table 6.1. Descriptive statistics for the effect of doses data.

	Mean	Median	SD	Skewness	Kurtosis	Min.	Max.
Treatment 1	2477.00	2481.00	483.47	-0.381	-1.434	1687.00	3020.00
Treatment 2	2075.00	1972.50	274.87	0.587	-1.546	1825.00	2527.00
Treatment 3	527.16	523.50	200.31	0.298	-1.579	317.00	842.00
Treatment 4	156.33	141.00	38.76	0.850	-1.030	127.00	227.00
Treatment 5	91.33	76.50	52.8	1.028	-0.588	44.00	193.00

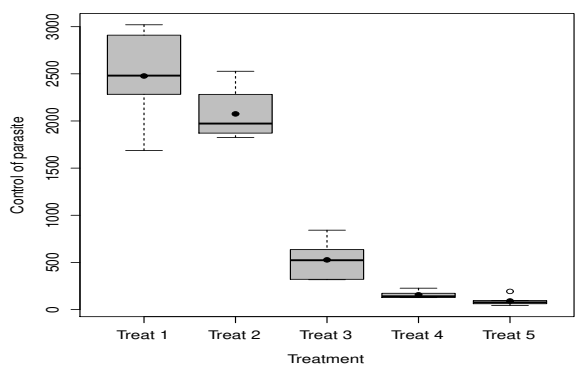

Figure 6.2. Boxplot for effect of doses data by treatments.

First, we present an exploratory and marginal analysis. In this way, we consider in this analysis the proposed GOLLMax2, GIG, GA and NO models. Figure 6.3a and 6.3b adjusted curves of the cumulative distribution and density functions, respectively, considered under the marginal analysis of the response variable y. As we can see due to the positive asymmetry, the normal model visually is not the most suitable.

Then, all variables involved in the study are: y_i: response variable control a parasite (log scale); d_{i1}: comparing Treatment 1 with Treatment 2; d_{i2}: comparing Treatment 1 with Treatment 3; d_{i3}: comparing Treatment 1 with Treatment 4; d_{i4}: comparing Treatment 1 with Treatment 5; for $i = 1, \ldots, 30$. We now present results for the GOLLMax2, GIG, GA and NO regression models by considering the following systematic structures as presented in (6.6):

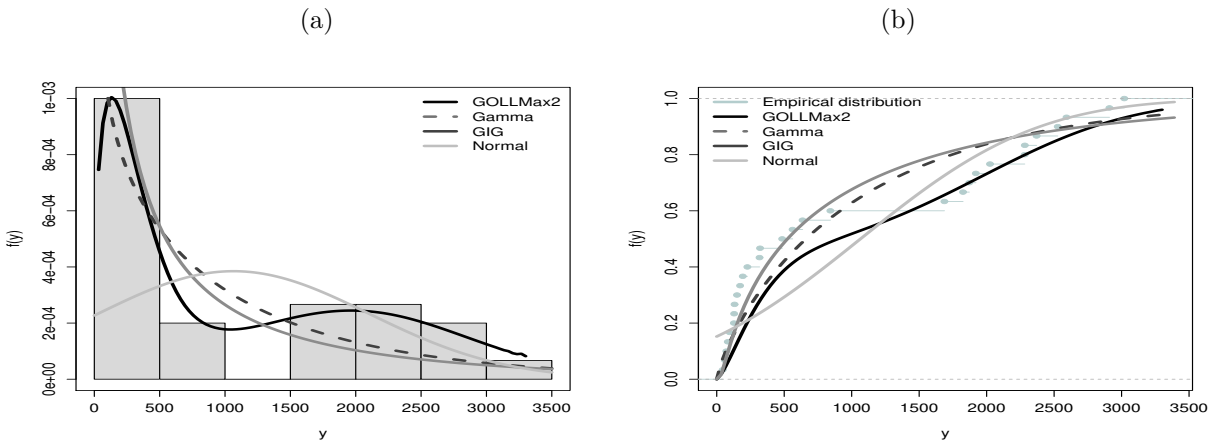

Figure 6.3. Plots for doses data data. (a) Histogram and estimated GOLLMax2, GIG, GA and NO densities. (b) Empirical distribution and estimated GOLLMax2, GIG, GA and NO distributions.

$$\mu_i = \exp(\psi_0 + \psi_1 d_{i1} + \psi_2 d_{i2} + \psi_3 d_{i3} + \psi_4 d_{i4}).$$

The values of the GD, AIC and BIC statistics of the fitted regression models are listed in Table 6.2. The estimates of the parameters, SEs and the associated p-values of the MLEs in Table 6.2. The figures in this table give evidence that the Treatment 1 and Treatment 2 considered control do not differ at a significance level of 5% for GOLLMax2, GIG and GA models. This fact confirms the exploratory analysis shown in Figure 6.2. The same is not true for the normal regression model.

Table 6.2. MLEs and information criteria for effect of doses data.

GOLLMax2				GIG			
$\boldsymbol{\theta}$	MLE	E.P	p-value	$\boldsymbol{\theta}$	MLE	E.P	p-value
ψ_0	7.808	0.102	<0.001	ψ_0	7.860	0.116	<0.001
ψ_1	-0.188	0.140	0.192	ψ_1	-0.153	0.162	0.354
ψ_2	-1.578	0.164	<0.001	ψ_2	-1.633	0.162	<0.001
ψ_3	-2.801	0.146	<0.001	ψ_3	-2.768	0.162	<0.001
ψ_4	-3.451	0.162	<0.001	ψ_4	-3.469	0.162	<0.001
$\log(\sigma)$	-3.181	0.650	-	$\log(\sigma)$	1.462	85.516	-
$\log(\nu)$	3.247	0.124	-	ν	-12.63	3.221	-
	GD	AIC	BIC		GD	AIC	BIC
	382.37	396.37	406.18		382.61	396.6	406.42
Gamma				Normal			
$\boldsymbol{\theta}$	MLE	E.P	p-value	$\boldsymbol{\theta}$	MLE	E.P	p-value
ψ_0	7.814	0.102	<0.001	ψ_0	7.814	0.040	<0.001
ψ_1	-0.177	0.170	0.31	ψ_1	-0.177	0.062	0.009
ψ_2	-1.547	0.170	<0.001	ψ_2	-1.547	0.192	<0.001
ψ_3	-2.762	0.170	<0.001	ψ_3	-2.762	0.635	0.0002
ψ_4	-3.300	0.170	<0.001	ψ_4	-3.300	1.086	0.005
$\log(\sigma)$	-1.218	0.127	-	$\log(\sigma)$	5.492	0.129	-
	GD	AIC	BIC		GD	AIC	BIC
	385.68	397.68	406.08		414.67	426.67	435.07

In Figure 6.4, we perform the residual analysis by plotting the quantile residuals rq_i (see subsection 6.3.1) against the index of observations for the fitted GOLLMax2 (Figure 6.4a), GIG (Figure

6.4b), GA (Figure 6.4c) and Normal (Figure 6.4d) regression models.

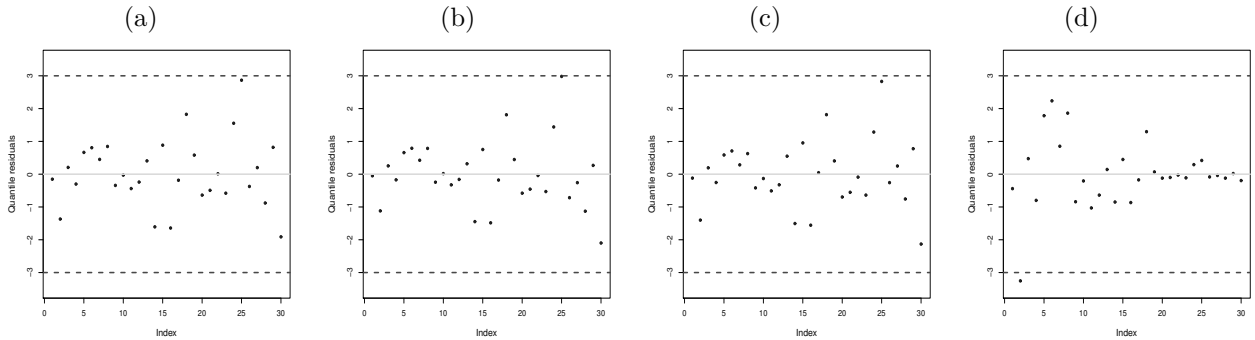

Figure 6.4. The index plot of the quantile residuals with range [-3, 3] for the fitted models to the effect of doses data. (a) GOLLMax2 regression model. (b) GA regression model. (c) GIG regression model. (c) NO regression model.

In Figure 6.5, gives the worm plot of the quantile residuals for the fitted GOLLMax2 (Figure 6.5a), GIG (Figure 6.5b), GA (Figure 6.5c) and Normal (Figure 6.5d) regression models. We conclude that Figures 5 and 8 support the hypothesis that the GOLLMax2 regression model is very suitable to fit these data

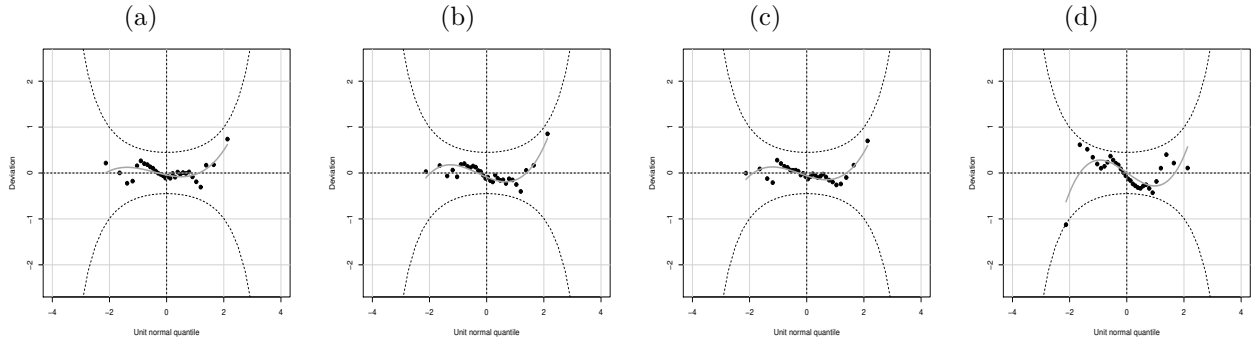

Figure 6.5. The worm plot of the quantile residuals for the fitted models to the effect of doses data. (a) GOLLMax2 regression model. (b) GA regression model. (c) GIG regression model. (c) NO regression model.

We compute case-deletion measures LD_i and GD_i defined in subsection 6.3.1. The results of such influence measure index plots are displayed in Figure 6.6a and Figure 6.6b, respectively. 6.6c shows the quality of the adjustment of the GOLLMax2 regression model by constructing the normal probability for $qr's$ for the waste diversion with simulated envelope. There is evidence of a good fit of the GOLLMax2 regression model.

6.5 Concluding Remarks

In general, in this application it appears that the proposed model GOLLMax2 can be an alternative in the analysis of experimental data. The GOLLMax2, GIG and GA models do not differ significantly from the measures GD, AIC and BIC. However, such models are not fitted. Regarding the diagnostic analysis, the proposed model presents behaviors similar to the models mentioned. Procedures

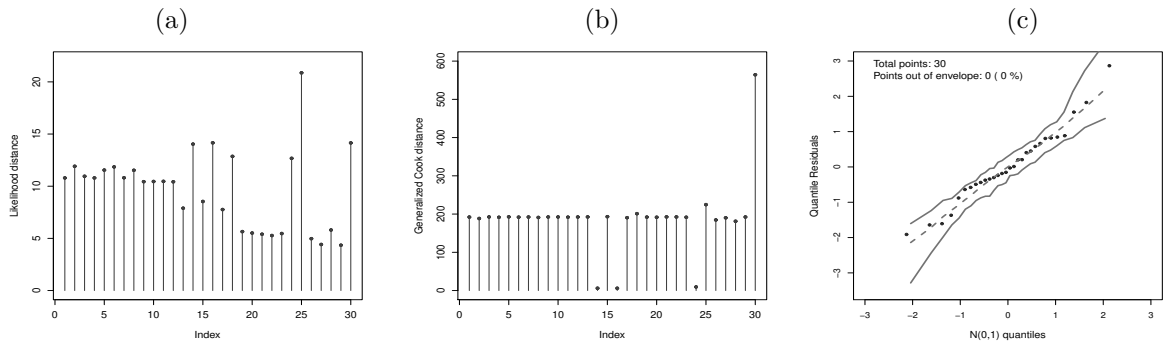

Figure 6.6. (a) Likelihood distance for for GOLLMax2 regression model to the effect dose data. (b) Generalized Cook distance for GOLLMax2 regression model to the effect dose data. (c) Normal probability plot for the $qr's$ with envelopes.

for fitting the GOLLMax2 median regression model and for model diagnostics are included in the `gamlss` package and available from the authors.

References

Atkinson, A.C. (1985). *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis*. Clarendon Press, Oxford.

Braga, A.S., Cordeiro, G.M., Ortega, E.M.M. and da Cruz, J.N. (2016). The odd log-logistic normal distribution: Theory and applications in analysis of experiments. *Journal of Statistical Theory and Practice*, **10**, 311-335.

Cordeiro, G.M., Ortega, E.M.M. and Silva, G.O. (2011). The exponentiated generalized gamma distribution with application to lifetime data. *Journal of Statistical Computation and Simulation*, **81**, 827-842.

Cordeiro, G.M., Alizadeh, M., Ozel, G., Hosseini, B., Ortega, E.M.M. and Altun, E. (2017). The generalized odd log-logistic family of distributions: properties, regression models and applications. *Journal of Statistical Computation and Simulation*, **87**, 908-932.

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

da Cruz, J.N., Ortega, E.M.M. and Cordeiro, G.M. (2016). The log-odd log-logistic Weibull regression model: Modelling, estimation, influence diagnostics and residual analysis. *Journal of Statistical Computation and Simulation*, **86**, 1516-1538.

Jørgensen, B. (1982). *The Theory of Dispersion Models*. Chapman and Hall: London, 1997.

Prataviera, F., Ortega, E.M., Cordeiro, G.M. and Braga, A. D. S. (2018). The heteroscedastic odd log-logistic generalized gamma regression model for censored data.*Communications in Statistics-Simulation and Computation*, 1-25.

Mudholkar, G.S. and Srivastava, D.K. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE transactions on reliability*, **42**, 299-302.

Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1-46.

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

7 A NEW DISTRIBUTION FOR RATES AND PROPORTIONS DATA

Abstract: We propose a new continuous distribution in the interval $(0, 1)$ based on the generalized odd log-logistic-G family, whose density function can be symmetrical, asymmetric, unimodal and bimodal. The new model is implemented using the `gamlss` packages in `R`. We obtain some of its structural properties. We propose an extended regression based on this distribution which includes as sub-models some important regressions. We employ a frequentist analysis and non-parametric and parametric bootstrap methods to estimate its parameters. Further, for different parameter settings and sample sizes, several simulations are conducted to study the empirical distribution of the maximum likelihood estimators. The empirical distribution of the quantile residuals is compared with the standard normal distribution. The extended regression is very useful for the analysis of real data and can give more realistic fits than other special regressions.
Keywords: Beta distribution; Generalized odd log-logistic; Maximum likelihood; Moment; Simulation.

7.1 Introduction

Many studies in several fields aim to determine how a set of explanatory variables influence other variables expressed as ratios or proportions, i.e., random experiments that produce results in the interval $(0, 1)$. Several researchers tried to model this type of data. For example, Ferrari and Cribari-Neto (2004) pioneered a regression in which the parameters are interpreted as mean and precision, Bayes *et al.* (2012) proposed a robust regression for proportions based on the beta rectangular distribution, Lemonte and Bazán (2016) defined a class of Johnson S_B distributions and its associated regression for rates and proportions, Mazucheli *et al.* (2019) proposed a unit-Lindley distribution and its associated regression for proportional data and Nakamura *et al.* (2019) defined a flexible distribution to deal with variables on the unit interval based on a transformation of the sinh-arcsinh distribution applied to modeling the points ratio of football teams.

It is common in practice to find proportional data with U and bimodal shapes. The known models applied in these situations generally have an initial structure based on the beta distribution, traditionally used to model constant, increasing, decreasing and unimodal data. For example, the *Human Development Index* (HDI) is adopted for comparison among countries, cities or other regions by measuring the degree of economic development and quality of life offered to the population. In this paper, we analyze the HDI of Brazilian cities in the States of Santa Catarina and Bahia. Figure 7.1 displays the histogram of these data ($n = 710$ observations) which refers to the first application, where a clear bimodal shape is evident.

Furthermore, the response variable in many practical situations can be related with covariates to explain the variability of proportional data. In these terms, our main aim is to propose a regression based on the *generalized odd log-logistic beta* ("GOLLBE" for short) distribution to model proportional data with bimodality.

The GOLLBE model is implemented numerically in the `gamlss` package of the `R` software to estimate the model parameters. The generalized additive model for location, scale and shape (GAMLSS) in `R` (Stasinopoulos and Rigby, 2007) is used to implement some new models. Here, we denote the package by `gamlss` and the modeling by GAMLSS. Recently, de Castro *et al.* (2010) described an application of the GAMLSS framework to fitting long-term survival models, Correa *et al.* (2012) implemented the

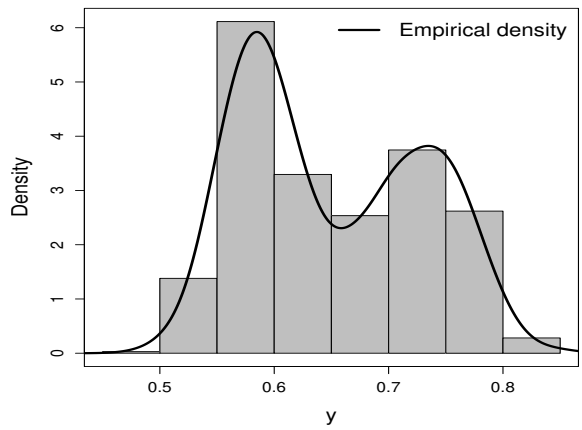

Figure 7.1. Histogram and plot of the empirical density for the HDI.

Kumaraswamy normal distribution in `gamlss` and make comparisons with Azzalini's skew normal distribution and Alizadeh *et al.* (2019) developed a new useful four-parameter extension of the Gumbel distribution with applications using `gamlss`. We also conduct hypothesis tests based on the asymptotic distribution of the maximum likelihood estimators (MLEs). Finally, through several simulation studies, we evaluate the performance of these estimators in the GOLLBE regression. Therefore, as an alternative to the maximum likelihood method, we also adopt the bootstrap method to estimate the model parameters.

In general, we have to verify the model assumptions after fitting a regression to a data set. If the model is not adequate, it can lead to misleading conclusions. It is also important to check for the presence of extreme or influential observations, which can cause distortions in the results. We adopt global influence measures based on case deletion to detect influential observations in the proposed regression. In order to check the model assumptions and detect atypical points, we consider the quantile residuals, and also perform Monte Carlo simulations to check the empirical distribution of these residuals using confidence bands constructed from generated envelopes (Atkinson, 1985).

The rest of the chapter is organized as follows. In Section 7.2, we introduce the GOLLBE distribution. Some structural properties of the GOLLBE distribution are addressed in 7.3 including ordinary and incomplete moments, generating function and mean deviations. In Section 7.4, we define a regression in the interval $(0, 1)$ and the classic inference and bootstrap method to estimate the model parameters. In Section 7.5, we perform and discuss some simulations. In Section 7.6, we consider diagnostics and residual analysis for the proposed regression. In Section 7.7, we provide two applications for the new regression. Section 7.8 offers some concluding remarks. Finally, we present some codes used in applications in Appendix C.

7.2 The GOLLBE distribution

A large number of distributions to extend well-known ones providing flexibility in modeling data has being investigated recently. In fact, Cordeiro *et al.* (2017) defined the cumulative distribution function (cdf) of the *generalized odd log-logistic-G* ("GOLL-G") family from a baseline cdf $G(y;\boldsymbol{\gamma})$ (depending on

a parameter vector $\boldsymbol{\gamma}$) by integrating the log-logistic density function, namely

$$F(y;\nu,\tau,\boldsymbol{\gamma}) = \int_0^{\frac{G(y;\boldsymbol{\gamma})^\nu}{1-G(y;\boldsymbol{\gamma})^\nu}} \frac{\tau\, w^{\tau-1}}{(1+w^\tau)^2} dw = \frac{G(y;\boldsymbol{\gamma})^{\nu\,\tau}}{G(y;\boldsymbol{\gamma})^{\nu\,\tau} + [1-G(y;\boldsymbol{\gamma})^\nu]^\tau}, \tag{7.1}$$

where $\nu > 0$ and $\tau > 0$ are two additional shape parameters.

Equation (7.1) is a wider continuous family including as special cases the odd log-logistic (Gleaton and Lynch, 2006) and proportional reversed hazard rate (Gupta and Gupta, 2007) distributions when $\nu = 1$ and $\tau = 1$, respectively. For $\nu = \tau = 1$, we obtain the baseline G distribution.

The probability density function (pdf) corresponding to (7.1) is

$$f(y;\nu,\tau,\boldsymbol{\gamma}) \quad = \quad \frac{\nu\,\tau\, g(y;\boldsymbol{\gamma}) G(y;\boldsymbol{\gamma})^{\nu\,\tau-1}\,[1-G(y;\boldsymbol{\gamma})^\nu]^{\tau-1}}{\{G(y;\boldsymbol{\gamma})^{\nu\,\tau} + [1-G(y;\boldsymbol{\gamma})^\nu]^\tau\}^2}, \tag{7.2}$$

where $g(y) = g(y;\boldsymbol{\gamma})$ is the baseline density. Further, we can omit the dependence on the vector $\boldsymbol{\gamma}$ and write simply $G(y) = G(y;\boldsymbol{\gamma})$. The density function (7.2) allows greater flexibility of its tails and can be widely applied in many areas of engineering and biology. It will be most tractable when $G(y)$ and $g(y)$ have closed-forms.

The interest in the beta distribution is due to its versatility to model random experiments that produce data in terms of proportions. The normal linear regression is widely used in empirical analysis, but such model becomes inappropriate in situations where the response variable represents rates and proportions. For this reason, we present a new extended beta distribution.

The pdf and cdf of the beta random variable W (for $0 < y < 1$), re-parametrized in terms of the mean parameter μ ($0 < \mu < 1$) and dispersion parameter σ ($0 < \sigma < 1$), are

$$g(y;\mu,\sigma) = \frac{\Gamma(\frac{1-\sigma^2}{\sigma^2})}{\Gamma\left(\frac{\mu(1-\sigma^2)}{\sigma^2}\right)\Gamma\left(\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)} y^{\frac{\mu(1-\sigma^2)}{\sigma^2}-1}(1-y)^{\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}} \tag{7.3}$$

and

$$G(y;\mu,\sigma) \quad = \quad I_y\left(\frac{\mu(1-\sigma^2)}{\sigma^2}, \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right) = \frac{B\left(y; \frac{\mu(1-\sigma^2)}{\sigma^2}, \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)}{B\left(\frac{\mu(1-\sigma^2)}{\sigma^2}, \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)}, \tag{7.4}$$

respectively, where $\Gamma(p) = \int_0^\infty w^{p-1}\,\mathrm{e}^{-w} dw$ (for $p > 0$) is the gamma function, $B(a,b) = \int_0^1 w^{a-1}\,(1-w)^{b-1}\,dw$ is the beta function (for $a > 0$ and $b > 0$) and $B(w;a,b) = \int_0^w t^{a-1}(1-t)^{b-1}\,dt$ is the incomplete beta function. We consider the notation $W \sim$BE(μ,σ). The mean and variance of W are $E(W) = \mu$ and $Var(W) = \sigma^2 V(\mu)$, where $V(\mu) = \mu(1-\mu)$ and then σ^2 is a multiplier factor in the variance of W.

By inserting (7.3) and (7.4) in Equation (7.2) with $\boldsymbol{\gamma} = (\mu,\sigma)$, the GOLLBE density function can be expressed as (for $0 < y < 1$)

$$\begin{aligned} f(y;\nu,\tau,\mu,\sigma) \quad = \quad & \frac{\nu\,\tau I_y\left(\frac{\mu(1-\sigma^2)}{\sigma^2}, \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)^{\nu\,\tau-1}\left[1-I_y\left(\frac{\mu(1-\sigma^2)}{\sigma^2}, \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)^\nu\right]^{\tau-1}}{\left\{I_y\left(\frac{\mu(1-\sigma^2)}{\sigma^2}, \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)^{\nu\,\tau} + \left[1-I_y\left(\frac{\mu(1-\sigma^2)}{\sigma^2}, \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)^\nu\right]^\tau\right\}^2} \times \\ & \frac{\Gamma(\frac{1-\sigma^2}{\sigma^2})}{\Gamma\left(\frac{\mu(1-\sigma^2)}{\sigma^2}\right)\Gamma\left(\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)} y^{\frac{\mu(1-\sigma^2)}{\sigma^2}-1}(1-y)^{\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}}, \end{aligned} \tag{7.5}$$

where $0 < \mu < 1$ is the mean parameter and $0 < \sigma < 1$ is the dispersion parameter (both refereeing to the baseline BE), and $\nu > 0$ and $\tau > 0$ are the shape parameters. We denote by $Y \sim$GOLLBE(μ,σ,ν,τ) a random variable with pdf (7.5). The GOLLBE distribution includes as special cases the following distributions:

- The new odd log-logistic beta (OLLBE) distribution when $\nu = 1$.
- The new exponentiated beta (EBE) distribution when $\tau = 1$.
- The beta (BE) distribution when $\tau = \nu = 1$.

Some possible shapes of the density (7.5) for selected parameter values, including well-known distributions, are displayed in Figure 7.2, thus emphasizing its bimodal shapes. Note that the additional shape parameters give flexibility to the new distribution. We have bimodality when $\tau \in (0, 1)$, among other shapes depending on the parameter values.

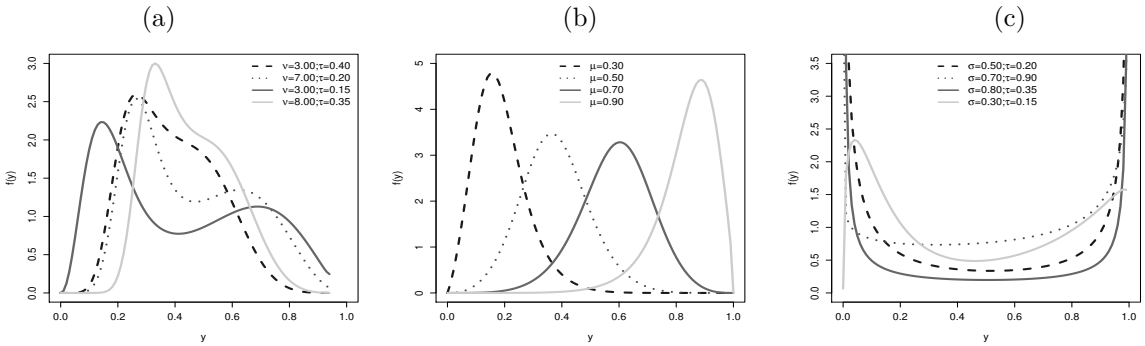

Figure 7.2. Plots of the GOLLBE density for selected parameter values. =(a) For $\mu = 0.3$ and $\sigma = 0.2$. (b) For $\sigma = 0.35$, $\nu = 0.5$ and $\tau = 2$. (c) For $\mu = 0.3$ and $\nu = 3$.=

The quantile function (qf) of Y has the form

$$y = Q(u) = Q_{\mathrm{BE}}\left(\left[\frac{\left(\frac{u}{1-u}\right)^{\frac{1}{\tau}}}{1+\left(\frac{u}{1-u}\right)^{\frac{1}{\tau}}}\right]^{\frac{1}{\nu}}; \mu, \sigma\right), \tag{7.6}$$

where $Q_{\mathrm{BE}}(u) = G^{-1}(u; \mu, \sigma)$ is the qf of the beta distribution.

The plots comparing the exact GOLLBE densities and histograms from three simulated data sets for some parameter values are displayed in Figure 7.3. These histograms are constructed based on 2,000 generated values.

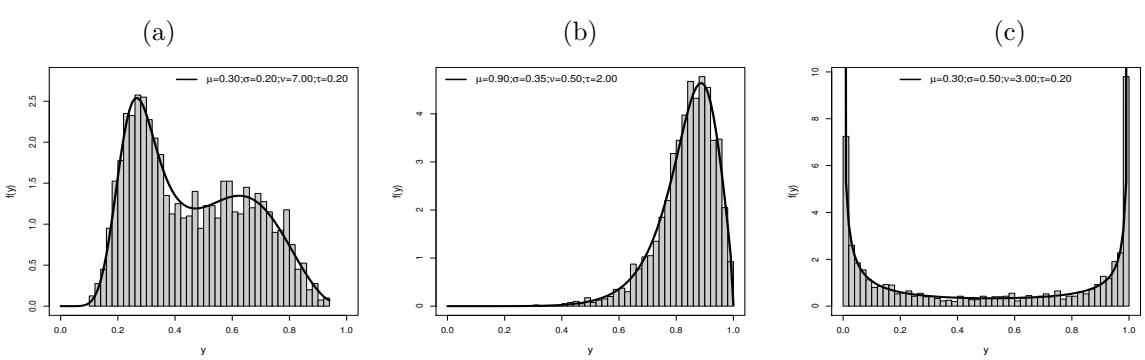

Figure 7.3. Histograms from generated GOLLBE densities for some shapes with adjusted density function (7.5). (a) Bi-modality. (b) Left skewness. (c) U-shape.

7.3 Structural properties

Linear representation

We provide a linear representation for the GOLLBE density. First, we define

$$F(y;\nu,\tau,\mu,\sigma) = \frac{z^{\nu\,\tau}}{z^{\nu\,\tau} + (1-z^{\nu})^{\tau}}, \tag{7.7}$$

where $z = z(y) = I_y\left(\mu\,(1-\sigma^2)/\sigma^2, (1-\mu)\,(1-\sigma^2)/\sigma^2\right) \in (0,1)$.

Second, we use a convergent power series for $z^{\nu\,\tau}$ (both parameters are positive real numbers)

$$z^{\nu\,\tau} = \sum_{k=0}^{\infty} a_k\; z^k, \tag{7.8}$$

where

$$a_k = a_k(\nu,\tau) = \sum_{j=k}^{\infty} (-1)^{k+j} \binom{\nu\,\tau}{j} \binom{j}{k}.$$

Third, we consider the generalized binomial expansion (under these parameters)

$$(1-z)^{\nu\,\tau} = \sum_{k=0}^{\infty} (-1)^k \binom{\nu\,\tau}{k} z^k. \tag{7.9}$$

Inserting (7.8) and (7.9) in Equation (7.7), we obtain

$$F(y;\nu,\tau,\mu,\sigma) = \frac{\sum_{k=0}^{\infty} a_k\, z^k}{\sum_{k=0}^{\infty} b_k\, z^k}, \tag{7.10}$$

where $b_k = b_k(\nu,\tau) = a_k(\nu,\tau) + (-1)^k \binom{\nu\,\tau}{k}$ (for $k \geq 0$). The ratio of the two power series in Equation (7.10) can be written as

$$F(y) = F(y;\nu,\tau,\mu,\sigma) = \sum_{k=0}^{\infty} c_k\, I_y\left(\mu\,(1-\sigma^2)/\sigma^2, (1-\mu)\,(1-\sigma^2)/\sigma^2\right)^k, \tag{7.11}$$

where $c_0 = a_0/b_0$ and the coefficients c_k's (for $k \geq 1$) are determined recursively from

$$c_k = c_k(\nu,\tau) = b_0^{-1}\left(a_k - \sum_{r=1}^{k} b_r\, c_{k-r}\right).$$

By differentiating (7.11), we can write the density, say $f(y) = f(y;\nu,\tau,\mu,\sigma)$, as

$$f(y) = c_1\; g(y;\mu,\sigma) + g(y;\mu,\sigma) \sum_{k=1}^{\infty} (k+1)\, c_{k+1}\, I_y\left(\mu\,(1-\sigma^2)/\sigma^2, (1-\mu)\,(1-\sigma^2)/\sigma^2\right))^k. \tag{7.12}$$

Fourth, the power series for the incomplete beta function ratio when $(1-\mu)(1-\sigma^2)/\sigma^2$ is real non-integer is

$$I_y\left(\mu\,(1-\sigma^2)/\sigma^2, (1-\mu)\,(1-\sigma^2)/\sigma^2\right) = \sum_{m=0}^{\infty} t_m\, y^{m+\mu\,(1-\sigma^2)/\sigma^2}, \tag{7.13}$$

where $t_m = t_m(\mu,(1-\sigma^2)/\sigma^2) = \frac{(1-(1-\mu)(1-\sigma^2)/\sigma^2)_m}{(\mu(1-\sigma^2)/\sigma^2+m)\,m!\,B(\mu\,(1-\sigma^2)/\sigma^2,(1-\mu)(1-\sigma^2)/\sigma^2)}$ (for $m \geq 0$), and $(p)_k = \Gamma(p+k)/\Gamma(p)$ is the ascending factorial.

Based on a result by Gradshteyn and Ryzhik (2000, 0.314) for a power series raised to an integer $k \geq 1$, we can write

$$\left(\sum_{m=0}^{\infty} t_m\, u^m\right)^k = \sum_{m=0}^{\infty} w_{k,m}\, u^m, \tag{7.14}$$

where the coefficients $w_{k,m} = w_{k,m}(\mu, (1-\sigma^2)/\sigma^2)$ (for $m = 1, 2, \ldots$) can be determined from the recurrence equation (with $w_{k,0} = t_0^k$)

$$w_{k,m} = (m\, t_0)^{-1} \sum_{r=1}^{m} [r(k+1) - m]\, t_r\, w_{k,m-r}.$$

Finally, combining (7.12), (7.13) and (7.14), we obtain

$$f(y) = c_1\; g(y; \mu(1-\sigma^2)/\sigma^2, (1-\mu)(1-\sigma^2)/\sigma^2) + g(y; \mu(1-\sigma^2)/\sigma^2, (1-\mu)(1-\sigma^2)/\sigma^2) \sum_{k=1}^{\infty} \sum_{m=0}^{\infty} (k+1)\, c_{k+1}\, w_{k,m}\, y^{m+k\mu\,(1-\sigma^2)/\sigma^2},$$

which can take the form

$$f(y) = c_1\; g(y; \mu(1-\sigma^2)/\sigma^2, (1-\mu)(1-\sigma^2)/\sigma^2) + \sum_{k=1}^{\infty} \sum_{m=0}^{\infty} s_{k,m}\, g(y; (k+1)\mu(1-\sigma^2)/\sigma^2 + m, (1-\mu)(1-\sigma^2)/\sigma^2), \tag{7.15}$$

where (for $k \geq 1$, $m \geq 0$)

$$s_{k,m} = s_{k,m}(\nu, \tau, \mu, \sigma) = \frac{(k+1)\,\Gamma((1-\sigma^2)/\sigma^2)\,\Gamma((k+1)\mu(1-\sigma^2)/\sigma^2 + m)}{\Gamma(\mu(1-\sigma^2)/\sigma^2)\,\Gamma((1+k\mu)(1-\sigma^2)/\sigma^2 + m)}\, c_{k+1}\, w_{k,m}.$$

Equation (7.15) reveals that the GOLLBE density is a linear combination of beta densities. Thus, several mathematical properties of the new distribution can be easily determined from those beta properties. For most applications, k and m could be limited to five.

Structural properties

Here, we obtain only two mathematical properties of $Y \sim$GOLLBE(μ, σ, ν, τ) from those of $W \sim$BE(μ, σ^2). First, the nth ordinary moment of Y can be found from (2.8) and the beta moments

$$\mu_n' = E(Y^n) = c_1\; \frac{B(n+\mu(1-\sigma^2)/\sigma^2, (1-\mu)(1-\sigma^2)/\sigma^2)}{B(\mu(1-\sigma^2)/\sigma^2, (1-\mu)(1-\sigma^2)/\sigma^2)} + \sum_{k=1}^{\infty} \sum_{m=0}^{\infty} s_{k,m}\, \frac{B(n+(k+1)\mu(1-\sigma^2)/\sigma^2 + m, (1-\mu)(1-\sigma^2)/\sigma^2)}{B(\mu(1-\sigma^2)/\sigma^2, (1-\mu)(1-\sigma^2)/\sigma^2)}.$$

The moment generating function (mgf) of Y follows from the mgf of $W \sim$BE(μ, σ) written in terms of Kummer's confluent hypergeometric function (of the first kind), namely

$$E(\mathrm{e}^{\mathrm{tW}}) = {}_1\mathrm{F}_1(\mu(1-\sigma^2)/\sigma^2; (1-\sigma^2)/\sigma^2) = \sum_{\mathrm{k}=0}^{\infty} \frac{(\mu(1-\sigma^2)/\sigma^2)_{\mathrm{k}}}{((1-\sigma^2)/\sigma^2)_{\mathrm{k}}\, \mathrm{k}!}.$$

We can write from (7.15) and the above result

$$E(\mathrm{e}^{\mathrm{tY}}) = \mathrm{c}_1\; {}_1\mathrm{F}_1(\mu(1-\sigma^2)/\sigma^2; (1-\sigma^2)/\sigma^2) + \sum_{\mathrm{k}=1}^{\infty} \sum_{\mathrm{m}=0}^{\infty} \mathrm{s}_{\mathrm{k,m}}\; {}_1\mathrm{F}_1((\mathrm{k}+1)\mu(1-\sigma^2)/\sigma^2 + \mathrm{m}; \mathrm{k}\mu(1-\sigma^2)/\sigma^2 + \mathrm{m} + (1-\sigma^2)/\sigma^2).$$

The plots below reveal the behavior of the expected value of the GOLLBE distribution and also the behavior of the sample mean ($\bar{Y}$) in relation to the parameter μ. In fact, Figure 7.4 shows the expected value of the new distribution versus one of the parameters for fixed values of the other parameters. Thus, we can note very different forms of $E(Y)$ depending on the selected parameters. Figure 7.5 shows that generating different samples under these scenarios, $\bar{Y}$ tends to follow a linear regression with the parameter μ. The simulations in Figure 7.5 are performed in the `R` software based on 300 replications.

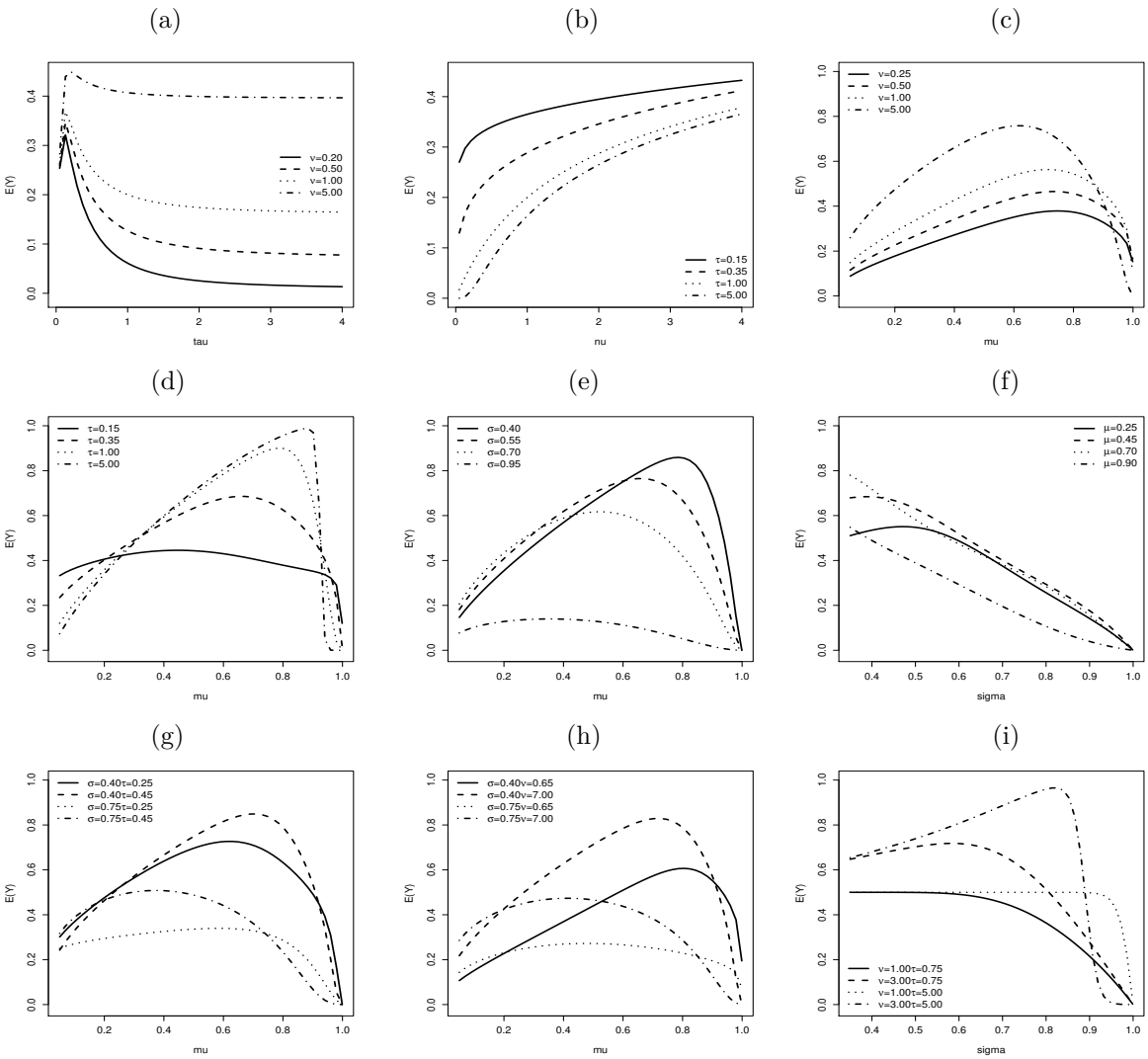

Figure 7.4. Typical forms for $E(Y)$ for selected parameter values. (a) Varying τ for $\mu = 0.2$, $\sigma = 0.4$ and ν fixed. (b) Varying ν for $\mu = 0.2$, $\sigma = 0.4$ and ν fixed. (c) Varying μ for $\sigma = 0.5$, $\tau = 0.45$ and ν fixed. (d) Varying μ for $\sigma = 0.45$, $\nu = 3.0$ and τ fixed. (e) Varying μ for $\nu = 3.0$, $\tau = 0.65$ and σ fixed. (f) Varying σ for $\nu = 8.0$, $\tau = 0.25$ and μ fixed. (g) Varying μ for $\nu = 7.0$, σ and τ fixed. (h) Varying μ for $\tau = 0.45$, σ and τ fixed. (i) Varying σ for $\mu = 0.5$, ν and τ fixed.

7.4 The GOLLBE regression and estimation

In several problems of the medical, biological, industrial and chemical areas, among others, it is of interest to verify if two or more variables are related in some way. When performing a regression analysis, one typically wishes to make inferences on the model parameters and use diagnostic tools to identify atypical observations.

Let $Y_1, \ldots, Y_n$ be independent random variables with each $Y_i \sim$GOLLBE$(\mu_i, \sigma, \nu, \tau)$ by assuming that the mean parameter μ_i of the baseline BE distribution varies across observations. We define the GOLLBE regression from the random component (7.5) and the systematic component

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \qquad i = 1, \ldots, n, \tag{7.16}$$

where $g : (0, 1) \to R$ is the link function, $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{ip})$ is a vector of known explanatory variables

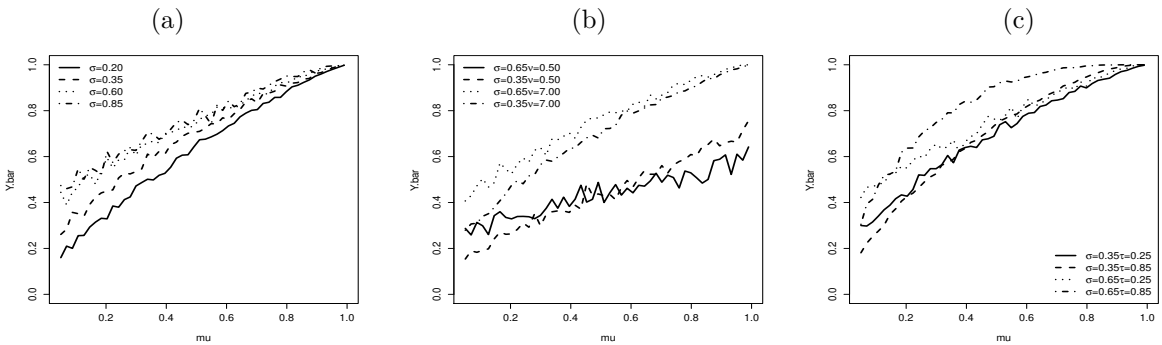

Figure 7.5. Plots of $\overline{Y}$ versus μ from simulated GOLLBE variates. (a) Varying μ for $\nu = 7.0$, $\tau = 0.25$ and σ fixed. (b) Varying μ for $\tau = 0.25$, σ and ν fixed. (c) Varying μ for $\nu = 7.0$, σ and τ fixed.

and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a vector of dimension p of unknown coefficients. We can write $g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ is a specified $n \times p$ matrix of full column rank $p < n$. The linear structure in $g(\boldsymbol{\mu})$ aims to explain the variability of the observations.

We consider that $g(\cdot)$ is a known one-to-one continuously twice differentiable function. We can choose different building functions such as the logit $g(\mu) = \log[\mu/(1-\mu)]$. The GOLLBE regression opens new possibilities for fitting many different types of proportional data, since the response distribution is much more flexible than the OLLBE, EBE and beta distributions.

Some works that may be developed in the future:

- Define a systematic component for the dispersion parameter σ similar to that one of the generalized linear models with dispersion covariates.
- Consider the GAMLSS approach, where all parameters of a distribution (not necessarily belonging to the exponential family) can be modeled by systematic parametric or semiparametric components.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ be a random sample of n independent observations, where the response variable y_i belongs to the interval $(0, 1)$. The log-likelihood function for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma, \nu, \tau)^\top$ can be expressed as

$$
\begin{aligned}
l(\boldsymbol{\theta}) \;=\;& n\,\log[\nu\,\tau\,\Gamma((1-\sigma^2)/\sigma^2)] + (\nu\,\tau - 1)\sum_{i=1}^{n}\log[h_i(y_i;\mu_i,\sigma)] - \\
& \sum_{i=1}^{n}\log\left[\Gamma(\mu_i(1-\sigma^2)/\sigma^2)\Gamma((1-\mu_i)(1-\sigma^2)/\sigma^2)\right] - \\
& 2\sum_{i=1}^{n}\log\left\{h_i^{\nu\,\tau}(y_i;\mu_i,\sigma) + [1 - h_i^{\nu}(y_i;\mu_i,\sigma)]^{\tau}\right\} + \\
& \sum_{i=1}^{n}(\mu_i(1-\sigma^2)/\sigma^2 - 1)\log(y_i) \\
& + \sum_{i=1}^{n}[(1-\mu_i)(1-\sigma^2)/\sigma^2 - 1]\log(1-y_i) + (\tau - 1)\sum_{i=1}^{n}\log[1 - h_i^{\nu}(y_i;\mu_i,\sigma)], \qquad (7.17)
\end{aligned}
$$

where $h_i(y_i;\mu_i,\sigma) = I_{y_i}\left(\mu_i\,(1-\sigma^2)/\sigma^2, (1-\mu_i)\,(1-\sigma^2)/\sigma^2\right)$. The MLEs of the unknown parameters are calculated by maximizing the log-likelihood function (7.17) with respect to $\boldsymbol{\theta}$. There is no closed-form expression for the MLE $\widehat{\boldsymbol{\theta}}$ but its computation can be performed using the `RS` maximization algorithm

(default in `gamlss` platform) in `R`; see Stasinopoulos and Rigby (2007) and Stasinopoulos *et al.* (2017). Initial values for $\boldsymbol{\beta}$ and σ are taken from the fit of the beta regression with $\nu = \tau = 1$. We can also consider the EBE regression for $\tau = 1$ and OLLBE regression for $\nu = 1$. The codes used in the applications are given in Appendix C.

Besides estimation of the model parameters, hypothesis tests can be performed based on the classical likelihood ratio (LR) statistics. For selection of the appropriate distribution, we use the global deviance (GD), say $GD = -2l(\hat{\boldsymbol{\theta}})$, where $l(\hat{\boldsymbol{\theta}})$ is the maximized log-likelihood function (7.17), and the generalized Akaike information criterion (GAIC) given by $GAIC(k) = GD + k \times df$, where df is the total degrees of freedom of the adjusted model and k is the penalty for each degree of freedom used. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are special cases of the GAIC(k) measure when $k = 2$ and $k = \log(n)$, respectively. We consider the GD, AIC and BIC criteria to select the best regressions.

7.4.1 Bootstrap re-sampling method

The bootstrap re-sampling method pioneered by Efron (1979) considers that the observed sample represents the population and it is widely used in different statistical situations. According to Moore (2006) the re-sampling methods allow to quantify the uncertainty by calculating the standard errors and confidence intervals as well as performing hypothesis tests. Based on the information obtained from one sample, B bootstrap samples of similar size to that of the observed sample are generated from which it is possible to estimate various characteristics of the population, such as the mean, variance, percentiles and so on. According to the literature, the re-sampling method may be non-parametric or parametric. In this study, we conduct non-parametric and parametric bootstrap methods. To perform the bootstrap procedure, we use the `boot` library available in `R`. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be an observed random sample and F the empirical distribution or the true distribution for the parametric case of $\mathbf{Y}$. A bootstrap sample $\mathbf{Y}^*$ is constructed by re-sampling nonparametric scheme with the replacement of n elements from the sample $\mathbf{Y}$. From B generated bootstrap samples, $Y_1^*, \ldots, Y_B^*$, the bootstrap replication of the parameter of interest for the bth sample is

$$\hat{\boldsymbol{\theta}}_b^* = s(Y_b^*),$$

which is the value of $\hat{\boldsymbol{\theta}}$ for sample T_b^* $(b = 1, \ldots, B)$.

The bootstrap estimator of the standard error (Efron and Tibshirani, 1993), say $\widehat{EP}_B$, is the standard deviation of the B parameter estimates computed from the bootstrap samples given by

$$\widehat{EP}_B = \left[\frac{1}{(B-1)} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \bar{\theta}_B \right)^2 \right]^{1/2},$$

where $\bar{\theta}_B = B^{-1} \sum_{b=1}^{B} \hat{\theta}_b^*$.

We describe the bias corrected and accelerated (BCa) method for constructing approximated confidence intervals based on the bootstrap re-sampling method.

BCa bootstrap interval

The bootstrap confidence intervals (CIs) based on the BCa method considers that the percentiles used in delimiting the CIs depend on the corrections to tendency $\hat{a}$ and acceleration $\hat{z}_0$. The bias-correction constant (Efron, 1987) $\hat{z}_0$ can be expressed as

$$\hat{z}_0 = \Phi^{-1} \left(Pr(\hat{\boldsymbol{\theta}}_b^* < \hat{\boldsymbol{\theta}}) \right), \quad b = 1, \ldots, B,$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution, $\Phi^{-1}(\cdot)$ is its inverse, $\hat{\boldsymbol{\theta}}$ is the MLE from the observed sample and $\hat{\boldsymbol{\theta}}_b^*$ is the MLE from the bth bootstrap sample.

Let $\hat{\boldsymbol{\theta}}_{(i)}$ be the MLE of $\boldsymbol{\theta}$ computed from the sample without the ith observation. Then, $\hat{a}$ has the form

$$\hat{a} = \frac{\sum_{i=1}^{n}\left[\hat{\boldsymbol{\theta}}_{(\cdot)} - \hat{\boldsymbol{\theta}}_{(i)}\right]^3}{6\left\{\sum_{i=1}^{n}\left[\hat{\boldsymbol{\theta}}_{(\cdot)} - \hat{\boldsymbol{\theta}}_{(i)}\right]^2\right\}^{3/2}}.$$

Note that $\hat{\boldsymbol{\theta}}_{(\cdot)} = \sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_{(i)}/n$.

The BCa bootstrap interval of coverage $100(1-2\alpha)\%$ is

$$\left[\hat{\boldsymbol{\theta}}^*_{(B\alpha_1)}, \hat{\boldsymbol{\theta}}^*_{(B\alpha_2)}\right],$$

where

$$\alpha_1 = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 + \Phi^{-1}(\alpha)}{1 - \hat{a}\left[\hat{z}_0 + \Phi^{-1}(\alpha)\right]}\right\} \quad \text{and} \quad \alpha_2 = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 + \Phi^{-1}(1-\alpha)}{1 - \hat{a}\left[\hat{z}_0 + \Phi^{-1}(1-\alpha)\right]}\right\}.$$

Here, α_1 and α_2 are corrections to the bootstrap percentiles and the other quantities were defined before.

7.5 Simulation study

We perform different simulation studies with and without covariates and a misspecification study for the GOLLBE regresssion. We calculate the MLEs $\hat{\mu}$, $\hat{\sigma}$, $\hat{\nu}$ and $\hat{\tau}$ for each replication and $n = 60, 250$ and 500. We repeat this process $1,000$ times and determine the average estimates (AEs), say $(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau})$, biases and means squared errors (MSEs). The generated variates from the GOLLBE distribution are obtained as follow: i) Generate $u \sim$Uniforme (0,1); ii) Calculate $y = Q(u)$ from equation (7.6) to generate values from the GOLLBE distribution.

We consider three scenarios by taking the shapes shown in Figure 7.3. In the first scenario (a), Figure 7.3a, we set $\mu = 0.30$, $\sigma = 0.20$, $\nu = 7.0$ and $\tau = 0.20$. In the second scenario (b), Figure 7.3b, we have $\mu = 0.90$, $\sigma = 0.35$, $\nu = 0.50$ and $\tau = 2.0$. In the third scenario (c), Figure 7.3c, we take $\mu = 0.30$, $\sigma = 0.50$, $\nu = 3.0$ and $\tau = 0.20$. The simulation results reported in Table 7.1 indicate that the MSEs of the MLEs of μ, σ, ν and τ decay toward zero when the sample size increases, as expected under first-order asymptotic theory.

7.5.1 Simulations for the GOLLBE regression

Various simulations are performed for the GOLLBE regression with three sample sizes ($n = 60, 250$ and 500). We explore two scenarios based on the systematic component (7.16). In the first scenario, $\beta_0 = 0.45$, $\beta_1 = -0.15$, $\sigma = 0.20$, $\nu = 0.45$, $\tau = 0.40$. In the second scenario, $\beta_0 = 0.10$, $\beta_1 = 0.85$, $\sigma = 0.20$, $\nu = 2.0$, $\tau = 0.95$. For both scenarios the systematic component is $\mu_i = \frac{\exp(\beta_0+\beta_1 x_{i1})}{1+\exp(\beta_0+\beta_1 x_{i1})}$ and $x_{i1} \sim \mathrm{N}(0.50, 0.65)$ for $i = 1, \ldots, n$.

The response variables $y_1, \ldots, y_n$ are generated from the GOLLBE regression in the following way:

i. For the first and second scenarios, generate $x_{i1} \sim$N(0.50, 0.65);

Table 7.1. AEs, Bias and MSEs for the estimates in the GOLLBE distribution under scenarios a, b and c.

scenario a									
	$n = 60$			$n = 250$			$n = 500$		
Parameters	AEs	Bias	MSE	AEs	Bias	MSE	AEs	Bias	MSE
μ	0.319	0.019	0.004	0.311	0.011	0.001	0.305	0.005	0.000
σ	0.201	0.001	0.002	0.197	-0.003	0.000	0.199	-0.001	0.000
ν	6.312	-0.688	10.803	6.522	-0.478	3.544	6.708	-0.292	1.541
τ	0.229	0.029	0.008	0.203	0.003	0.001	0.203	0.003	0.001
scenario b									
	$n = 60$			$n = 250$			$n = 500$		
Parameters	AEs	Bias	MSEs	AEs	Bias	MSEs	AEs	Bias	MSEs
μ	0.897	-0.003	0.000	0.898	-0.002	0.000	0.898	-0.002	0.000
σ	0.312	-0.038	0.011	0.337	-0.013	0.004	0.342	-0.008	0.002
ν	0.533	0.033	0.142	0.510	-0.010	0.007	0.505	0.005	0.000
τ	1.801	-0.199	0.547	1.925	-0.075	0.292	1.954	-0.046	0.156
scenario c									
	$n = 60$			$n = 250$			$n = 500$		
Parameters	AEs	Bias	MSEs	AEs	Bias	MSEs	AEs	Bias	MSEs
μ	0.310	0.010	0.007	0.319	0.019	0.005	0.307	0.007	0.001
σ	0.502	0.002	0.007	0.491	-0.009	0.002	0.501	0.001	0.001
ν	2.944	-0.056	0.384	2.891	-0.109	0.000	2.937	-0.063	0.104
τ	0.229	0.029	0.018	0.229	0.000	0.001	0.205	0.005	0.001

ii. For both scenarios, calculate μ_i from the above systematic component;

iii. Generate $u_i \sim$ U(0,1);

iv. Use steps i., ii. and iii. to calculate the simulated observations y_i's from (7.6).

The simulation results reported in Tables 7.2 and 7.3 indicate that the AEs of the parameters tend to be closer to the true parameter values when the sample size n increases in agreement with the law of large numbers.

Table 7.2. The AEs, biases and MSEs based on 1,000 simulations for the GOLLBE regression under scenario 1 with $\beta_{10} = 0.45$, $\beta_{11} = -0.15$, $\sigma = 0.20$, $\nu = 0.45$ and $\tau = 0.40$.

	$n = 60$			$n = 250$			$n = 500$		
$\boldsymbol{\theta}$	AEs	Bias	MSEs	AEs	Bias	MSEs	AEs	Bias	MSEs
β_0	0.442	-0.008	0.162	0.452	0.002	0.045	0.453	0.003	0.025
β_1	-0.150	0.000	0.049	-0.149	0.001	0.007	-0.147	-0.003	0.004
σ	0.188	-0.012	0.004	0.195	-0.005	0.001	0.196	-0.004	0.000
ν	0.602	0.152	0.498	0.479	0.029	0.045	0.461	0.011	0.020
τ	0.395	-0.005	0.029	0.398	-0.002	0.008	0.394	-0.006	0.004

7.5.2 Misspecification Study

We investigate the behavior of the MLEs of the parameters in the GOLLBE regression when it is poorly specified by carrying out Monte Carlo simulations based on 1,000 replications. The response values are generated from the BE distribution by taking $\beta_0 = 0.35$, $\beta_1 = 0.15$ and $\sigma = 0.3$. We consider the generalized beta (GBE) density function

$$f(y;\mu,\sigma,\nu,\tau) = \frac{\tau\nu^{\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}} y^{\frac{\tau\,\mu(1-\sigma^2)}{\sigma^2}-1}(1-y^{\tau})^{\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}-1}}{B\left(\frac{\mu(1-\sigma^2)}{\sigma^2},\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)\left[\nu+(1-\nu)y^{\tau}\right]^{\frac{(1-\sigma^2)}{\sigma^2}}}, \quad 0 < y < 1,$$

Table 7.3. The AEs, biases and MSEs based on 1,000 simulations for the GOLLBE regression under scenario 2 with $\beta_{10} = 0.10$, $\beta_{11} = 0.85$, $\sigma = 0.20$, $\nu = 2.0$ and $\tau = 0.95$.

	$n = 60$			$n = 250$			$n = 500$		
$\boldsymbol{\theta}$	AEs	Bias	MSEs	AEs	Bias	MSEs	AEs	Bias	MSEs
β_0	0.125	0.025	0.054	0.096	-0.004	0.016	0.100	0.000	0.008
β_1	0.866	0.016	0.009	0.851	0.001	0.002	0.851	0.001	0.001
σ	0.173	-0.027	0.005	0.197	-0.003	0.002	0.198	-0.002	0.001
ν	2.938	0.938	17.024	2.214	0.214	1.005	2.081	0.081	0.322
τ	0.847	-0.103	0.201	0.941	0.063	0.063	0.944	-0.006	0.032

where $0 < \mu < 1$, $0 < \sigma < 1$, $\nu > 0$ and $\tau > 0$. The observations are simulated by taking $\beta_0 = 0.35$, $\beta_1 = 0.15$, $\sigma = 0.3$, $\nu = 2.0$ and $\tau = 0.85$.

The logit-normal (LOGITNO) density function is defined by $(0 < y < 1)$

$$f(y;\mu,\sigma) = \frac{1}{y(1-y)\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log\left(\frac{y}{1-y}\right) - \log\left(\frac{\mu}{1-\mu}\right)}{\sigma}\right)^2\right\},$$

where $0 < \mu < 1$ and $\sigma > 0$. The parameters are fixed at $\beta_0 = 0.35$, $\beta_1 = 0.15$ and $\sigma = 0.3$.

We fit all these regressions under the systematic component (given in Section 7.4) with $x_{i1} \sim \mathrm{N}(0.50, 0.65)$ (for $i = 1, \ldots, 60$) to each generated data set using the `gamlss` packages. The results are listed in Tables 7.4, 7.5 and 7.6. In addition to the AEs, biases and MSEs, we present the mean measures of GD, AIC and BIC. They indicate that there are small sample biases in the parameter estimation. The average measures of GD, AIC and BIC for the fitted GOLLBE regression are very close to those values obtained from the true regressions used in the generation of the response values. Hence, the GOLLBE regression provides consistent MLEs even when the data are generated from different regressions.

Table 7.4. The AEs, biases and MSEs for the GOLLBE regression based on 1,000 simulations of the beta regression for $n = 60$ with $\beta_0 = 0.35$ and $\beta_1 = 0.15$.

	BE			GOLLBE		
$\boldsymbol{\theta}$	AEs	Bias	MSEs	AEs	Bias	MSEs
β_0	0.352	0.002	0.012	0.273	-0.077	0.121
β_1	0.150	0.000	0.018	0.154	0.004	0.022
σ	0.294	-0.006	0.003	0.249	-0.051	0.011
ν				1.720	-0.280	8.007
τ				0.795	-0.055	0.196
	GD=-65.26	AIC=-59.26	BIC=-52.97	GD=-66.89	AIC=-56.89	BIC=-46.41

Table 7.5. The AEs, biases and MSEs for the GOLLBE regression model based on 1,000 simulations of the GBE regression for $n = 60$ with $\beta_0 = 0.35$ and $\beta_1 = 0.15$.

	GBE			GOLLBE		
$\boldsymbol{\theta}$	AEs	Bias	MSEs	AEs	Bias	MSEs
β_0	0.638	0.288	0.126	0.874	0.524	0.392
β_1	0.152	0.002	0.019	0.159	0.009	0.023
σ	0.286	-0.014	0.003	0.235	-0.065	0.015
ν	1.593	-0.407	2.681	1.196	-0.804	3.703
τ	0.989	0.139	0.189	0.887	0.037	0.325
	GD=-82.61	AIC=-72.61	BIC=-62.14	GD=-83.69	AIC=-73.69	BIC=-63.22

Table 7.6. The AEs, biases and MSEs for the GOLLBE regression based on 1,000 simulations of the LOGITNO regression for $n = 60$ with $\beta_0 = 0.35$ and $\beta_1 = 0.15$.

	LOGITNO			GOLLBE		
$\boldsymbol{\theta}$	AEs	Bias	MSEs	AEs	Bias	MSEs
β_0	0.352	0.002	0.003	0.318	-0.032	0.028
β_1	0.146	-0.004	0.005	0.145	-0.005	0.005
σ	0.294	-0.006	0.001	0.116	-0.184	0.036
ν				1.509	-0.491	3.023
τ				0.786	-0.064	0.172
	GD=-151.83	AIC=-145.83	BIC=-139.54	GD=-153.54	AIC=-143.54	BIC=-133.07

7.6 Diagnostics and residual analysis

An important step after the estimation of the parameters is the diagnosis of abnormalities of the fitted regression, for example, detect influential observations that may cause significant distortions in the analysis results. This step is known as sensitivity analysis. A first tool to perform this analysis, as stated before, is by means of global influence starting from case-deletion. Case-deletion is a common approach to study the effect of dropping the ith case from a data set. A global influence measure as a generalization of the Cook distance (Xie and Wei, 2007) is the standardized norm of $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$ given by

$$GD_i = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^\top \big[\ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}})\big](\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}), \tag{7.18}$$

where $-\ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}})$ is the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$.

Another measure to evaluate the influence of an observation (Cook and Weisberg, 1982), called the log-likelihood distance, is the difference between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{(i)}$ on the log-likelihood scale, namely

$$LD_i = 2\Big[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)})\Big], \tag{7.19}$$

where $l(\hat{\boldsymbol{\theta}})$ is the maximized log-likelihood for the full sample and $l(\hat{\boldsymbol{\theta}}_{(i)})$ is the maximized log-likelihood for the sample excluding the ith observation.

In the analysis of a fitted regression we have to check possible deviations from the model assumptions. In this context, we aim to detect the presence of outliers in the data set and evaluate their impact on the inferential results. Therefore, an analysis of the residuals can help to validate the robustness of the inferential results. We consider the *quantile residuals* and prove from the simulations that they have some interesting properties in the GOLLBE regression under different systematic components.

The quantile residuals (Dunn and Smyth, 1996) for the GOLLBE regression have the form

$$\widehat{rq}_i = \Phi^{-1}\left\{\frac{I_{y_i}^{\hat{\nu}\,\hat{\tau}}\left(\hat{\mu}_i\,(1-\hat{\sigma}^2)/\hat{\sigma}^2, (1-\hat{\mu}_i)\,(1-\hat{\sigma}^2)/\hat{\sigma}^2\right)}{I_{y_i}^{\hat{\nu}\,\hat{\tau}}\left(\hat{\mu}_i\,(1-\hat{\sigma}^2)/\hat{\sigma}^2, (1-\hat{\mu}_i)\,(1-\hat{\sigma}^2)/\hat{\sigma}^2\right) + \left[1 - I_{y_i}^{\hat{\nu}}\left(\hat{\mu}_i\,(1-\hat{\sigma}^2)/\hat{\sigma}^2, (1-\hat{\mu}_i)\,(1-\hat{\sigma}^2)/\hat{\sigma}^2\right)\right]^{\hat{\tau}}}\right\}, \tag{7.20}$$

where $I_{y_i}(\cdot)$ is defined in equation (7.4) and $\Phi^{-1}(\cdot)$ is the standard normal qf.

Atkinson (1985) suggested the construction of envelopes to provide better interpretation of the probability normal plot of the residuals. These envelopes are simulated confidence bands that contain the residuals such that the majority of points will be within these bands and randomly distributed if the regression is well-fitted.

We also use Worm Plots (WP) as the technique to check the adjustment quality. Worm plots of the residuals were introduced by Buuren and Fredriks (2001). The general idea of these plots is to identify regions (intervals) of an explanatory variable within which the model does not fit adequately to the data. The WP is a detrended normal QQ-plot of the residuals. Model inadequacy is indicated when

many observations lie outside the point-wise 95% confidence bands or when the points follow a systematic shape. For example, the interpretations of the shapes of the WP are: a vertical shift, a slope, a parabola or a S shape, thus indicating a misfit in the mean, variance, skewness and excess kurtosis of the residuals, respectively.

7.7 Applications

In this section, we provide two applications to real data to illustrate the flexibility of the GOLLBE regression. The calculations are performed using the `gamlss` packages in the `R` software. In the first application, we present a real situation in which the behavior of the data is bimodal. In the second application, we consider a regression with systematic component (7.16) and the parametric and non-parametric bootstrap methods. The proposed regression is compared to other nested regressions. We also consider the GBE distribution as a competitive model.

7.7.1 Municipal HDI in Brazil

Our aim here is to prove empirically the potentiality of the proposed distribution to model data that present a complex shape including bimodality. The Human Development Index (HDI) is a measure of long-term progress in three basic dimensions of human development which takes into account education, income and longevity indexes. The HDI is the geometric mean of the normalized indexes for the three dimensions of human development. The current data set with $n = 710$ observations is available at the link *http://atlasbrasil.org.br/2013/*, where the observations were measured during the 2010 census. We analyze a municipal HDI (MHDI) data set in the cities of the States of Santa Catarina and Bahia (Brazil).

Tables 7.7 and 7.8 give the MLEs of the parameters, the values of the GD, AIC and BIC criteria and LR tests for the extra parameters ν and τ in the proposed distribution, respectively. We note that the lowest values of these criteria refer to the GOLLBE distribution, which can be chosen in this case.

Table 7.7. MLEs of the model parameters and statistical measures for HDI data.

Model	logit(μ)	logit(σ)	log(ν)	log(τ)	GD	AIC	BIC
GOLLBE	0.357	-2.509	2.149	-1.809	-1807.03	-1799.03	-1780.70
	(0.001)	(0.003)	(0.045)	(0.034)			
OLLBE	0.672	-2.780	-	-1.451	-1740.11	-1734.11	-1720.41
	(0.005)	(0.012)	-	(0.062)			
EBE	-0.190	-0.991	2.135	-	-1607.52	-1601.52	-1587.82
	(0.012)	(0.017)	(0.044)	-			
BE	0.623	-1.606	-	-	-1585.00	-1581.00	-1571.87
	(0.013)	(0.030)	-	-			
GBE	2.530	-1.701	-4.410	1.475	-1627.29	-1619.29	-1601.03
	(0.734)	(0.396)	(0.861)	(0.190)			

For the LR tests, we reject the three null distributions in favor of the GOLLBE distribution. As an alternative to the goodness-of-fit criteria, we display the plots of the fitted GOLLBE, OLLBE, EBE, GBE and BE densities and the histogram of the data in Figure 7.6a. The plots of the estimated cumulative distributions and empirical cdf are given in Figure 7.6b.

Two plots of the ordered quantile residuals versus the standard normal quantiles (QQ-plots) are displayed in Figure 7.7 to verify the adequacy of the fitted GOLLBE and BE regressions. We note that

Table 7.8. LR tests for HDI data.

Models	Hypotheses	Statistic w	p-value
GOLLBE vs OLLBE	$H_0 : \nu = 1$ vs $H_1 : H_0$ is false	67.69	<<0.0001
GOLLBE vs EBE	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	194.37	<<0.0001
GOLLBE vs BE	$H_0 : \nu = \tau = 1$ vs $H_1 : H_0$ is false	222.37	<<0.0001

(a) (b)

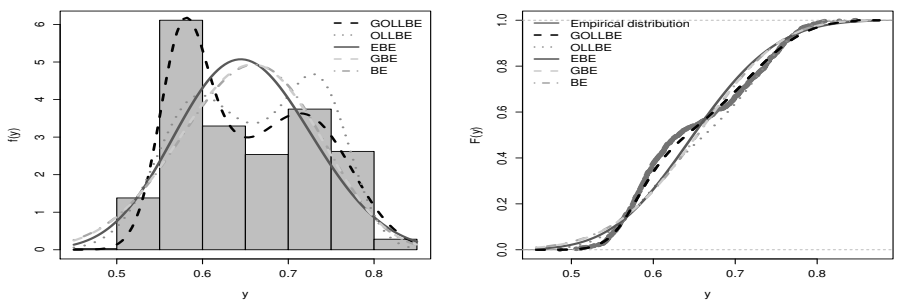

Figure 7.6. Some plots for HDI data. (a) Histogram and estimated GOLLBE, OLLBE, EBE, BE and GBE densities. (b) Estimated GOLLBE, OLLBE, EBE, BE and GBE cumulative distributions and empirical distribution.

the quantile residuals follow more approximately a normal distribution for the GOLLBE distribution. In fact, these plots reveal that the new distribution provides a good fit to the MHDI data.

(a) (b)

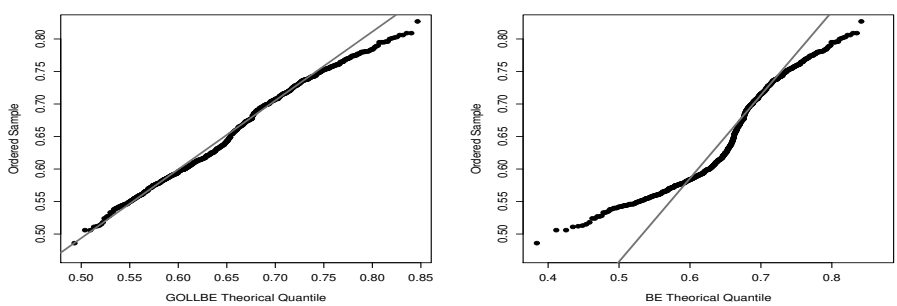

Figure 7.7. QQ-plots for MHDI data. (a) For the GOLLBE distribution. (b) For the BE distribution.

7.7.2 Body fat percentages in Australia

In the second application, we consider the Australian Institute of Sport (AIS) data set included in the library `sn` in the `R` software with $n = 202$ observations. The variables are: the body fat percentage (y_i = Bfat), weight (in kg) (x_{i1} = wt) and (x_{i2} = sex) a factor with levels (f: female and m: male) of each athlete (for $i = 1, \ldots, 202$). In the second part of this application, we also fit the GOLLBE regression to y with the explanatory variables x_1 and x_2 and present a diagnostic analysis based on the quantile residuals.

First, we present a marginal analysis considering only the response variable y. In Table 7.9, we give the MLEs of the parameters for some distributions and the values of the GD, AIC and BIC statistics. It is clear that the smallest values of these statistics are associated with the GOLLBE distribution, which

can be chosen in this case. The LR statistics for the extra parameters ν and τ in this distribution are listed in Table 7.10.

Table 7.9. MLEs in some fitted distributions for AIS data and statistical measures.

Model	logit(μ)	logit(σ)	log(ν)	log(τ)	GD	AIC	BIC
GOLLBE	-2.332	-2.481	1.990	-1.625	-633.27	-625.27	-612.04
OLLBE	-1.859	-2.477	-	-1.093	-613.81	-607.81	-597.89
EBE	-2.854	-1.203	1.876	-	-594.54	-588.54	-578.62
BE	-1.855	-1.563	-	-	-589.32	-585.32	-578.70
GBE	-1.471	-1.397	-0.436	0.015	-590.37	-582.37	-569.14

The LR values support the rejection of the null models in relation to the proposed distribution.

Table 7.10. LR tests for AIS data.

Models	Hypotheses	Statistic w	p-value
GOLLBE vs OLLBE	$H_0 : \nu = 1$ vs $H_1 : H_0$ is false	19.46	<0.0001
GOLLBE vs EBE	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	38.73	<0.0001
GOLLBE vs BE	$H_0 : \nu = \tau = 1$ vs $H_1 : H_0$ is false	43.95	<0.0001

The histogram of the current data set and the plots of the estimated GOLLBE, OLLBE, EBE, GBE and BE densities are displayed in Figure 7.8a. The plots of the corresponding estimated cumulative and empirical distributions are given in Figure 7.8b. These plots reveal that both GOLLBE and OLLBE models provide good fits to the current data.

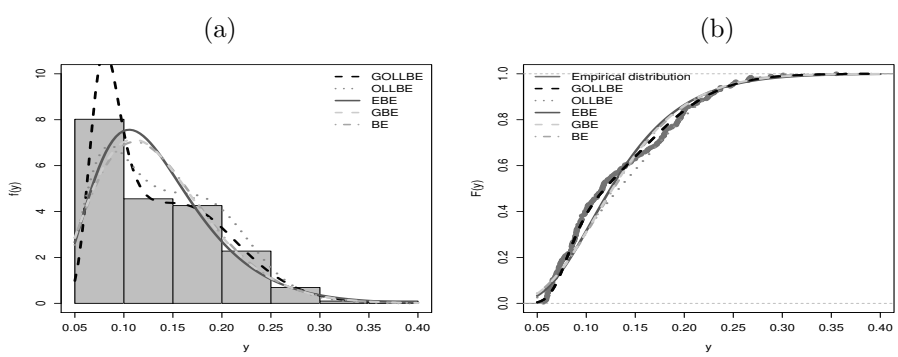

Figure 7.8. Plots for AIS data. (a) Histogram and estimated GOLLBE, OLLBE, EBE, BE and GBE densities. (b) Estimated GOLLBE, OLLBE, EBE, BE and GBE cumulative and empirical distributions.

Next, we fit some regressions to the body fat percentages with the systematic component

$$\mu_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})}, \qquad i = 1, \ldots, 202.$$

Table 7.11 provides the MLEs, SEs and p-values obtained from the fitted GOLLBE and BE regressions to the AIS data. The figures in this table reveal that the explanatory variables x_1 and x_2 are significant (at the 5% level) in modeling the body fat percentages.

The GD, AIC and BIC values for the fitted GOLLBE, OLLBE, EBE, BE and GBE regressions are listed in Table 7.12. We conclude that the GOLLBE and EBE regressions are better that the other competitive regressions independently of the criteria and then they can chosen to explain these data.

Table 7.11. MLEs, SEs and p-values for the GOLLBE and BE regression models fitted for AIS data.

GOLLBE				BE			
Parameter	Estimate	SE	p-Value	Parameter	Estimate	SE	p-Value
β_0	-3.444	0.385	< 0.0001	β_0	-2.955	0.112	< 0.0001
β_1	0.023	0.002	< 0.0001	β_1	0.020	0.001	< 0.0001
β_2	-1.203	0.124	< 0.0001	β_2	-1.073	0.047	< 0.0001
logit(σ)	-2.188	0.340	-	logit(σ)	-2.306	0.054	-
log(ν)	1.480	0.837	-				
log(τ)	-0.231	0.345	-				

Table 7.12. Statistical measures for five fitted regressions to the AIS data.

Model	GD	AIC	BIC
GOLLBE	-855.40	-843.40	-823.55
OLLBE	-851.73	-841.73	-825.19
EBE	-855.10	-845.10	-828.55
BE	-851.09	-843.09	-829.85
GBE	-854.94	-842.94	-823.09

Since the GOLLBE regression is the most suitable model for the AIS data, its parameters are also estimated by the non-parametric and parametric bootstrap methods. The bootstrap procedure is performed using the `boot` library in the `R` software with $B = 1,000$. We obtain the AEs computed from B bootstrap samples and the BCa CIs as described in Section 7.4.1. The AEs, their SEs and CIs are given in Table 7.13. The figures in Tables 7.11 and 7.13 indicate that the non-parametric bootstrap method is more efficient than the other two methods.

Table 7.13. AEs and CIs based on the non-parametric and parametric bootstrap re-sampling methods for the AIS data under the GOLLBE regression.

Non-parametric				Parametric			
Parameter	Average	SE	95% CIs - BCa	Parameter	Average	SE	95% CIs - BCa
β_0	-3.097	0.124	(-3.346; -2.860)	β_0	-3.344	0.242	(-4.352; -3.207)
β_1	0.021	0.001	(0.018; 0.024)	β_1	0.022	0.002	(0.019; 0.029)
β_2	-1.104	0.050	(-1.206; -1.013)	β_2	-1.178	0.089	(-1.507; -1.106)
logit(σ)	-2.462	0.136	(-2.694; -2.178)	logit(σ)	-2.284	0.257	(-2.393; -1.464)
log(ν)	0.581	0.185	(0.272; 0.998)	log(ν)	1.309	0.507	(1.099; 3.424)
log(τ)	-0.335	0.198	(-0.708, 0.044)	log(τ)	-0.314	0.290	(-0.524; 0.655)

The case deletion measures $GD_i(\boldsymbol{\theta})$ and $LD_i(\boldsymbol{\theta})$ given by (7.18) and (7.19) for the AIS data are displayed in Figure 7.9, respectively, where the cases ♯56 and ♯121 are possible influential observations.

Figure 7.10 displays the quantile residuals calculated from equation (7.20) for the chosen GOLLBE regression. In Figure 7.10a, we plot the quantile residuals against the fitted values. In Figure 7.10b, we plot these residuals against the index. The simulated envelop is displayed in Figure 7.10c. These plots reveal that the quantile residuals are arranged randomly in the interval $(-3, 3)$ and that they are within the confidence bands. Based on these analyses, there is a strong evidence of a good fit of the GOLLBE regression to the body fat data.

The worm plots obtained from all fitted regressions are displayed in Figure 7.11. We note that the GOLLBE and EBE regressions do not present any trend (vertical shift, slope, quadratic or cubic shapes). However, for the OLLBE, GBE and BE regressions, the normalized quantile residuals exhibit a quadratic or U shapes, thus indicating a problem with their skewness. Hence, based on the plots in

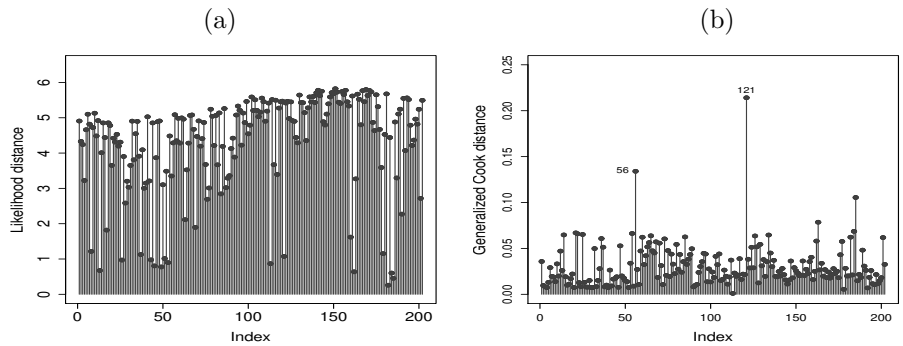

Figure 7.9. Plots for global influence. (a) Likelihood distance against index. (b) Generalized Cook's distance against index.

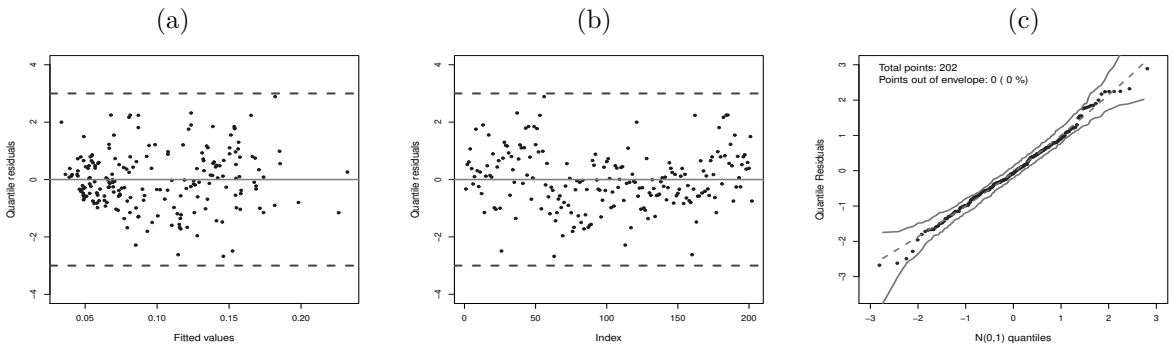

Figure 7.10. The quantile residuals from the GOLLBE regression. (a) Residuals against fitted values. (b) Residuals against index. (c) Simulated envelope for the residuals.

Figure 7.11 and the AIC and BIC measures in Table 7.12, we can conclude that the GOLLBE and EBE regressions are the most suitable models in relation to the others to explain AIS data.

In Figure 7.12, we present the predicted curves for both regressions. These plots are constructed using the function *update()* in the `R` software. It is evident that the proposed regression can be an interesting alternative in terms of prediction. We confirm the figures in Table 7.11, i.e. two distinct clusters in relation to the explanatory variable x_2. The percentages y are always higher for females than for males.

We can note from Table 9 that the covariables x_1 and x_2 are significant at a 5% significance level. The percentage of body fat tends to increase when the weight of individuals increases. Further, there is a significant difference between men and women regarding the percentage of body fat.

7.8 Conclusions

We propose a new continuous distribution with four parameters, called the generalized odd log-logistic beta (GOLLBE) distribution, for bimodal data in the unit interval. Its main advantage is the flexibility in accommodating different forms of the density function, such as U, unimodal and bimodal. We also investigate some f its structural properties. We define a new regression based on the GOLLBE distribution to the analysis of proportions and estimate its parameters by maximum likelihood. The bootstrap method is considered as an alternative for obtaining parameter estimates and confidence intervals. In addition, we perform diagnostic and residual analysis to verify the regression assumptions.

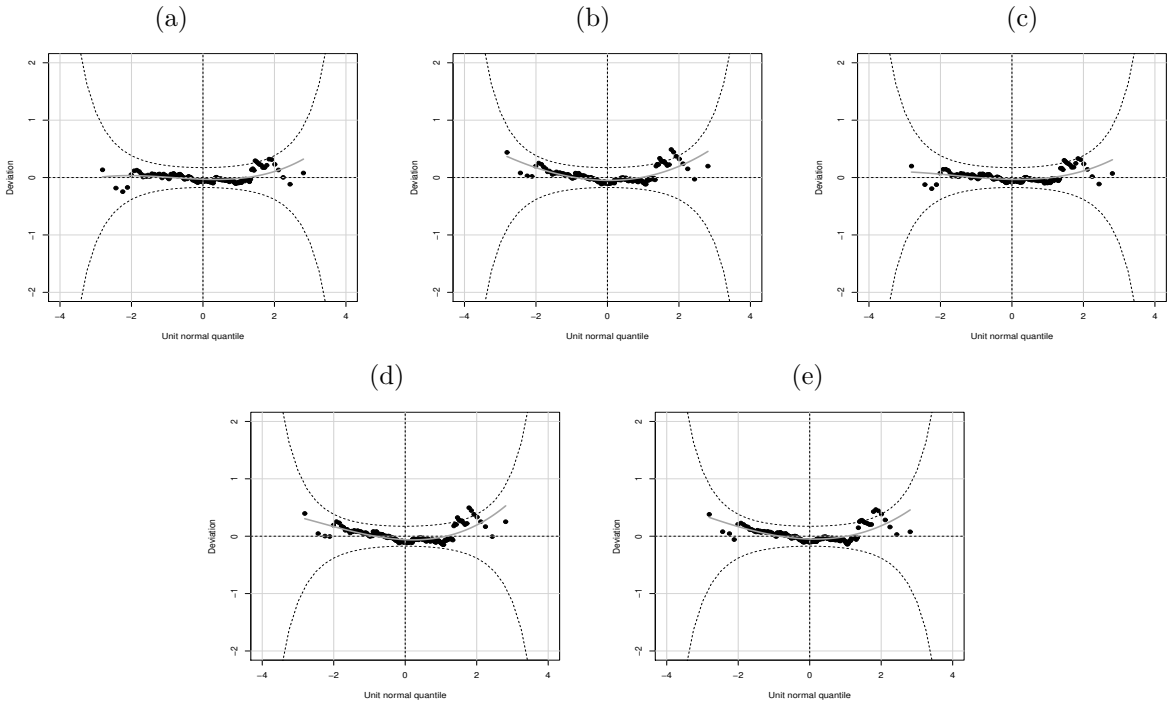

Figure 7.11. Worm plots in some fitted regressions for AIS data. (a) GOLLBE. (b) OLLBE. (c) EBE. (d) BE. (e) GBE.

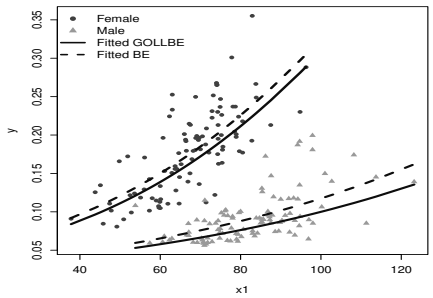

Figure 7.12. Plots of AIS data for y against x_1 (for each level of x_2) with fitted lines for the GOLLBE and BE regressions.

The usefulness of the new distribution and regression is illustrated by means of two real data sets, thus showing that the estimation methods present good results and that the residuals and sensitivity analysis can be helpful in choosing an appropriate model. Future work may be developed to deal with other estimation methods, such as Bayesian approach, and some studies can be conducted to verify the model robustness with respect to outliers. Further, zero-inflated regression models can be developed. Also, the GAMLSS models, parametric and/or semi-parametric models with random effects can be investigated.

References

Alizadeh, M., Ramires, T. G., MirMostafaee, S. M. T. K., Samizadeh, M. and Ortega, E. M. (2019). A new useful four-parameter extension of the Gumbel distribution: Properties, regression model and

applications using the GAMLSS framework. *Communications in Statistics-Simulation and Computation*, **48**, 1746-1767.

Atkinson, A. C. (1985). *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. University Press: Oxford.

Bayes, C.L., Bazán, J.L. and García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis*, **4**, 841-866.

Buuren, S.V. and Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, **20**, 1259-1277.

Cordeiro, G.M., Alizadeh, M., Ozel, G., Hosseini, B., Ortega, E.M.M. and Altun, E. (2017). The generalized odd log-logistic family of distributions: properties, regression models and applications. *Journal of Statistical Computation and Simulation*, **87**, 908-932.

Correa, M. A., Nogueira, D. A. and Ferreira, E. B. (2012). Kumaraswamy normal and Azzalini's skew normal modeling asymmetry. *Sigmae*, **1**, 65-83.

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

De Castro, M., Cancho, V. G. and Rodrigues, J. (2010). A hands-on approach for fitting long-term survival models under the GAMLSS framework. *Computer Methods and Programs in Biomedicine*, **97**, 168-177.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1-26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall: New York.

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

Ferrari, S.L.P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799-815.

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

Gradshteyn, I.S. and Ryzhik, I.M. (2000). *Table of Integrals, Series and Products*. Academic Press, San Diego.

Gupta, R.C. and Gupta, R.D. (2007). Proportional reversed hazard rate model and its applications. *Journal of Statistical Planning and Inference*, **137**, 3525-3536.

Lemonte, A.J. and Bazán, J.L. (2016). New class of Johnson distributions and its associated regression model for rates and proportions. *Biometrical Journal*, **58**, 727-746.

Mazucheli, J., Menezes, A.F.B. and Chakraborty, S. (2019). On the one parameter unit-Lindley distribution and its associated regression model for proportion data. *Journal of Applied Statistics*, **46**, 700-714.

Nakamura, L. R., Cerqueira, P.H.R., Ramires, T.G., Pescim, R.R., Rugny, R A. and Stasinopoulos, D.M. (2019). A new continuous distribution on the unit interval applied to modelling the points ratio of football teams. *Journal of Applied Statistics*, **46**, 416-431.

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

Stasinopoulos, D.M., Rigby, R.A., Heller, G.Z., Voudouris V. and De Bastiani, F. (2017), *Flexible Regression and Smoothing: Using GAMLSS in R*, Chapman and Hall/CRC, New York.

Xie, F.C. and Wei, B.C. (2007). Diagnostics analysis in censored generalized Poisson regression model. *Journal of Statistical Computation and Simulation*, **77**, 695-708.

8 CONCLUSION

In this work we propose parametric and semiparametric regression models based on the family *generalized odd log-logistic-G*. Thus, two new models were proposed for the *generalized odd log-logistic Maxwell* (GOLLMax) distribution for for data on positive support and the *generalized odd log-logistic beta* (GOLLBE) distribution for data analysis in the unit interval. Various mathematical properties of the GOLLMax and GOLLBE distribution are investigated. We show that it can accommodate various shapes of the skewness, kurtosis and bi-modality. The former class of GOLLMax regression models is very suitable for modeling censored and uncensored lifetime data.

Based on the GOLLMax distribution, we propose a reparametrization of the model in terms of median, denoted by GOLLMax2. This chapter needs further studies on the reparametrized model. For example, to verify the properties of the model, simulation studies should be developed. Applications in different areas must be considered, as well as censored and uncensored data can be an alternative work.

Based on the GOLLMax distribution, we propose a new distribution called *zero adjusted generalized odd log-logistic Maxwell* (ZAGOLLMax). For this model, we present a ZAGOLLMax semiparametric regression model to analyze soil microbiology data. We also propose a mixture model called *generalized odd log-logistic Maxwell mixture* (GOLLMaxM), with application to a prostate cancer dataset.

Considering data in the unit interval, the GOLLBE distribution was proposed. This model presents more flexible shapes in against to the beta distribution, which is considered as the base line distribution. Such flexibility can be verified in the shapes that the probability density function assumes, as bimodality, right asymmetry, left asymmetry, U shape. The model proves to be a very interesting alternative, especially when it is considered a regression structure.

We use the `gamlss` script in the `R` package to obtain the maximum likelihood estimates and perform asymptotic tests for the model parameters based on the asymptotic distribution of the estimates.

As future perspective of work, the following possibilities can be highlighted: based on the GOLLMax2 model, consider the parametric and semiparametric regression models for censored and uncensored data, regression model with cure rate, inflated of zeros model and competitive risks. For the GOLLBE model, parametric and semiparametric regression models can be considered, considering model with inflated data of zeros, inflated of ones and or inflated of zeros and ones. In addition to further studies on the robustness of the model in a regression structure when in the presence of autlier observations

APPENDICES

Appendix A: Codes for Chapter 2

Here, we present the codes implemented in the GAMLSS package in the software R. The pdf, cdf, qf and the samples generator functions are

```
library(flexsurv); library(numDeriv); library(gamlss);
library(gamlss.cens);  #required packages source('https:xxxxx')
#implemented codes dGOLLMax(x,mu,sigma,nu,tau) #pdf
pGOLLMax(x,mu,sigma,nu,tau) #cdf
qGOLLMax(u,mu,sigma,nu,tau) #qf
rGOLLMax(n,mu,sigma,nu,tau) #samples generator
```

Example of the code used to obtain the estimates of the semiparametric regression model in the application 1.

```
fit1 = gamlss(y~cs(x1,df=5),sigma.formula =~1,nu.formula =~1,
family = GOLLMax(),n.cyc=200,c.crit=0.01,
data=data1)
```

Example of the code used to obtain the estimates of the semiparametric regression model in the application 2.

```
fit2 = gamlss(Surv(y,delta)~cs(x1,df=5)+x2+x5,sigma.formula
=~1,nu.formula =~1, family = cens(GOLLMax()),n.cyc=200,c.crit=0.01,
data=data2)
```

Appendix B: Codes for Chapter 3

Here, we present the codes implemented in the gamlss package in the software R. The pdf, cdf, qf and the samples generator functions are

```
library(gamlss); library(flexsurv); library(numDeriv) #required
packages source('https:xxxxx') #implemented codes
dZAGOLLMax(x,mu,sigma,nu,tau) #pdf
pZAGOLLMax(x,mu,sigma,nu,tau) #cdf
qZAGOLLMax(u,mu,sigma,nu,tau) #qf
rZAGOLLMax(n,mu,sigma,nu,tau) #samples generator
```

Example of the code used to obtain the estimates of semiparametric regression model.

```
fit = gamlss(y~cs(x1,df=3)+cs(x2,df=3)+cs(x3,df=3)+Block +
Trat,sigma.formula =~1,nu.formula =~1,tau.formula
=~cs(x1,df=3)+cs(x2,df=3)+cs(x3,df=3)+Block + Trat, family =
ZAGOLLMax(),n.cyc=300,c.crit=0.01, data=microbiological.data)
```

Appendix C: Codes for Chapter 3

The proposed GOLLBE model was implemented in the `gamlss` packages in the `R` software. The implemented structure and the codes of the applications are available for access via Github at the links described below.

The file with the implemented structure can be accessed by `source('https://gist.githubusercontent.com/fabiopviera/69847c04cf02c3f4df078bd748dfb6f4/raw/ec07e163314eb55d8063b9d03c467f0782003ceb/GOLLBE.r')` or `https://gist.github.com/fabiopviera`. The pdf, cdf, qf and the samples generator functions are:

```
library(flexsurv); library(numDeriv); library(gamlss);

#implemented code
dGOLLBE(x,mu,sigma,nu,tau) #pdf
pGOLLBE(x,mu,sigma,nu,tau) #cdf
qGOLLBE(u,mu,sigma,nu,tau) #qf
rGOLLBE(n,mu,sigma,nu,tau) #samples generator
```

Example of the script used to obtain the estimates for Application 1, can be accessed by `https://gist.github.com/fabiopviera/da2a74f5306f167dda12ee086b96cb8b`.

```
fit_BE<-gamlss(y~1,family=BE,n.cyc=300,c.crit=0.01)
fit_EBE<-gamlss(y~1,family=GOLLBE,n.cyc=300,c.crit=0.01,tau.start=1,tau.fix=T,
mu.start=fit_BE$mu.fv,sigma.start=fit_BE$sigma.fv)
fit_OLLBE<-gamlss(y~1,family=GOLLBE,n.cyc=300,c.crit=0.01,nu.start=1,nu.fix=T,
mu.start=fit_EBE$mu.fv,sigma.start=fit_EBE$sigma.fv,tau.start=1)
fit_GOLLBE<-gamlss(y~1,family=GOLLBE,n.cyc=300,c.crit=0.01,mu.start=fit_BE$mu.fv,
sigma.start=fit_BE$sigma.fv,nu.start=fit_EBE$nu.fv)
```

Example of the script used to obtain the estimates for Application 2, can be accessed by `https://gist.github.com/fabiopviera/94fad80f93907782b0da83e98ae03ebd`.

```
fit_BE<-gamlss(y~x1+x2,family=BE)
fit_EBE<-gamlss(y~x1+x2,family=GOLLBE,n.cyc=200,c.crit=0.01,
tau.start=1,tau.fix=T,mu.start=fit_BE$mu.fv,sigma.start=fit_BE$sigma.fv)
fit_OLLBE<-gamlss(y~x1+x2,family=GOLLBE,n.cyc=200,c.crit=0.01,
nu.start=1,nu.fix=T,mu.start=fit_BE$mu.fv,sigma.start=fit_BE$sigma.fv)
fit_GOLLBE<-gamlss(y~x1+x2,family=GOLLBE,n.cyc=200,c.crit=0.01,
sigma.start=fit_BE$sigma.fv,nu.start=fit_EBE$nu.fv)
```