ATLANTIS PRESS
Journal of Statistical Theory and Applications, Vol. 16, No. 4 (December 2017) 547–564

The Log-gamma-logistic Regression Model: Estimation, Sensibility and Residual Analysis

Elizabeth M. Hashimoto
Departamento Acadêmico de Matemática,
Universidade Tecnológica Federal do Paraná
Londrina, Brazil
ehashimoto@utfpr.edu.br

Edwin M.M. Ortega*
Departamento de Ciências Exatas,
Universidade de São Paulo
Piracicaba, Brazil
edwin@usp.br

Gauss M. Cordeiro
Departamento de Estatística,
Universidade Federal de Pernambuco,
Recife, Brazil
gausscordeiro@gmail.com

G.G. Hamedani
Department of Mathematics, Statistics and Computer Science,
Marquette University, Milwaukee, USA
gholamhoss.hamedani@marquette.edu

Received 27 February 2016

Accepted 18 February 2017

In this paper, we formulate and develop a log-linear model using a new distribution called the log-gamma-logistic. We show that the new regression model can be applied to censored data since it represents a parametric family of models that includes as sub-models several widely-known regression models and therefore can be used more effectively in the analysis of survival data. We obtain maximum likelihood estimates of the model parameters by considering censored data and evaluate local influence on the estimates of the parameters by taking different perturbation schemes. Some global-influence measurements are also investigated. Further, for different parameter settings, sample sizes and censoring percentages, various simulations are performed. In addition, the empirical distributions of some modified residuals are displayed and compared with the standard normal distribution. These studies suggest that the residual analysis usually performed in normal linear regression models can be extended to modified deviance residuals in the proposed regression model applied to censored data. We demonstrate that our extended regression model is very useful to the analysis of real data and may give more realistic fits than other special regression models.

Keywords: Censored data; gamma-log-logistic distribution; regression model; residual analysis; sensitivity analysis.

*Corresponding author.

1. Introduction

The log-logistic distribution is widely used in survival analysis and it is an alternative to the Weibull and lognormal distributions since it presents a hazard rate function (hrf) that increases, reaches a peak after some finite period and then declines gradually. The properties of the log-logistic distribution make it an attractive alternative to the log-normal and Weibull distributions in the analysis of survival data (Collet, 2003). Recently, for any continuous baseline G distribution, Zografos and Balakrishnan (2009) proposed a generalized gamma-generated distribution with an extra positive parameter. A three-parameter continuous model called the gamma-log-logistic (GLL) distribution, which extends the log-logistic model, investigated by Ramos et al. (2013).

One way to study the effects of the explanatory variables on the lifetime or survival time is through a regression location-scale model, also known as the accelerated lifetime model. These models consider that the response variable belongs to a family of distributions characterized by a parameter location and scale parameter. However, for the case of parametric models, it is considered that the lifetime follows a continuous distribution. Regression models can be proposed in different forms in survival analysis. Among them, the location-scale regression model (Lawless, 2003) is distinguished and it is frequently used in clinical trials. So, in the context of survival analysis, some distributions have been used as an alternative to analyze censored data.

In this paper, we define a location-scale regression model using the GLL distribuition called the *log-gamma-logistic* (LGL) regression model. We consider a classic analysis for the LGL regression model. The inferential part was carried out using the asymptotic distribution of the maximum likelihood estimators (MLEs), which in situations when the sample size is small, may present difficult results to be justified. As an alternative to classic analysis, we explore the use of bootstrap method for survival time analysis as a feasible alternative. After modeling, it is important to check some model assumptions as well as to conduct a robustness study in order to detect influential or extreme observations that can cause distortions in the results of the analysis. We develops a similar methodology to detect influential subjects in LGL regression models with censored data.

We compare two residuals to assess departures from the error assumptions and to detect outlying observations in LGL regression models with censored observations. For different parameter settings, sample sizes and censoring percentages, various simulation studies are performed and the empirical distribution of each residual is displayed and compared with the standard normal distribution.

The rest of the paper is outlined as follows. In Section 2, we consider a brief study on the GLL distribution and present certain the characterizations of LGL distribution. In Section 3, we define a LGL regression model of location-scale form, in addition to the MLEs, bootstrap method and the results from various simulation studies are explored. In Section 4, we use several diagnostic measures considering three perturbation schemes, case-deletion and the generalized leverage in LGL regression models with censored observations. Section 5 deals with the definition and discussion of the residuals and presents and comments the results from various simulation studies. In Section 6, a real data set is analyzed and some conclusion are addressed in Section 7.

2. The log-gamma-logistic distribution

Many papers have been published on the log-logistic (LL) distribution, such as: Shoukri et al. (1988), who studied this distribution to describe phenomena involving precipitation in Canada; Shayan et al. (2011), who proposed LL regression with four parameters that allows description of failure rate data with bathtub distribution; Santana et al. (2012), who presented a generalization of

the LL distribution and its properties; and Zare et al. (2012), who used the LL distribution to analyze data on breast cancer patients in the southern region of Iran. The probability density function (pdf) and cumulative distribution function (cdf) of the LL distribution are given by

$$g(t) = \frac{\alpha}{\lambda^{\alpha}} t^{\alpha-1} \left[1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right]^{-2} \quad \text{and} \quad G(t) = 1 - \left[1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right]^{-1}, \tag{2.1}$$

respectively, where $\lambda > 0$ is the scale parameter and $\alpha > 0$ is the shape parameter. From the works cited above, it can be noted that the LL distribution has been widely used in data analysis to describe natural phenomena and to predict future events, as is the case of water movement in the oceans. However, in some situations, the observed data have uncommon behavior, so the LL distribution is not suitable to describe and predict the phenomenon of interest. Recent developments have been made to define new generated families to control skewness and kurtosis through the tail weights and provide great flexibility in modeling skewed data in practice, the generators pioneered by Eugene et al. (2002), Cordeiro and de Castro (2011) and Alexander et al. (2012).

Recently, Zografos and Balakrishnan (2009) and Ristic and Balakrishnan (2012) proposed a family of univariate distributions generated by gamma random variables. For any baseline cdf $G(t)$, they defined the *gamma-G distribution* with pdf $f(t)$ and cdf $F(t)$ by

$$f(t) = \frac{1}{\Gamma(\phi)} \{-\log[1 - G(t)]\}^{\phi-1} g(t) \tag{2.2}$$

and

$$F(t) = \frac{\gamma(\phi, -\log[1 - G(t)])}{\Gamma(\phi)} = \frac{1}{\Gamma(\phi)} \int_0^{-\log[1-G(t)]} u^{\phi-1} \mathrm{e}^{-u} du,$$

respectively, for $\phi > 0$, where $g(t) = dG(t)/dt$, $\Gamma(a) = \int_0^{\infty} u^{a-1} \mathrm{e}^{-u} du$ is the gamma function, and $\gamma(a,z) = \int_0^z u^{a-1} \mathrm{e}^{-u} du$ is the incomplete gamma function. The gamma-G distribution has the same parameters of the G distribution plus an additional shape parameter $\phi > 0$. Each new gamma-G distribution can be obtained from a specified G distribution. For $\phi = 1$, the G distribution is a basic exemplar of the gamma-G distribution with a continuous crossover towards cases with different shapes (for example, a particular combination of skewness and kurtosis).

Thus, the GLL distribution, proposed by Ramos et al. (2013) is obtained by inserting (2.1) in equation (2.2). So, we have

$$f(t) = \frac{\alpha}{\lambda^{\alpha}\Gamma(\phi)} t^{\alpha-1} \left[1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right]^{-2} \left\{\log\left[1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right]\right\}^{\phi-1}, \quad t > 0,$$

where $\phi > 0$ and $\alpha > 0$ are shape parameters and $\lambda > 0$ is a scale parameter.

For $\phi = 1$, we obtain the LL distribution. Like the LL and GLL distributions can be employed in engineering and can be used to model reliability in survival problems. Plots of the GLL density function for selected parameter values are displayed in Figure 1a. Plots of the GLL hrf for some parameter values are displayed in Figure 1b. The new hrf can have four types of shapes: increasing, decreasing, bathtub and unimodal.

To obtain a distribution that belongs to the location-scale class of models, we consider the transformation of the random variable $Y = \log(T)$. Thus, setting $\alpha = 1/\sigma$ and $\lambda = \exp(\mu)$, the pdf of Y

(a) (b)

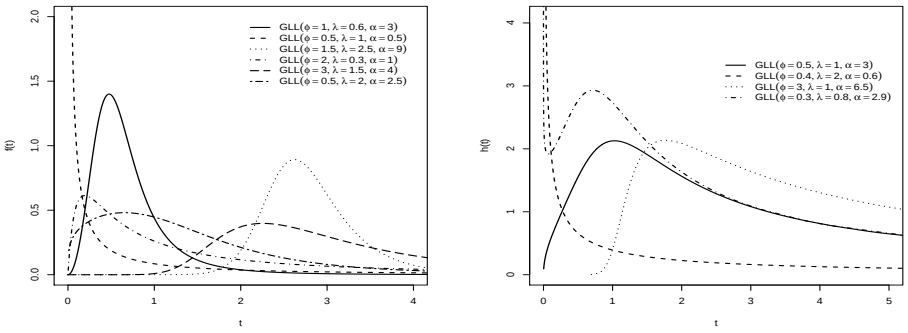

Fig. 1. (a) Plots of the GLL pdf for some parameter values. (b) Plots of the GLL hrf for some parameter values.

is given by

$$f(y) = \frac{1}{\sigma\Gamma(\phi)} \exp\left(\frac{y-\mu}{\sigma}\right) \left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-2} \left\{\log\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\phi-1}, \quad -\infty < y < \infty, \tag{2.3}$$

where $\phi > 0$, $\sigma > 0$ and $-\infty < \mu < \infty$. Therefore, Y is a random variable having the *log-gamma logistic* (LGL) distribution. Plots of the pdf (2.3) for selected parameter values are given in Figure 2. These plots show great flexibility for different values of the shape parameter ϕ with $\mu = 0$ and $\sigma = 1$. Thus,

$$\text{if} \quad T \sim \text{GLL}(\phi, \lambda, \alpha) \quad \text{then} \quad Y = \log(T) \sim \text{LGL}(\phi, \mu, \sigma).$$

Moreover, the survival function corresponding to (2.3) is

$$S(y) = 1 - \frac{1}{\Gamma(\phi)} \gamma\left(\log\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right], \phi\right).$$

On the other hand, we define the standardized random variable $Z = (Y-\mu)/\sigma$ with density function

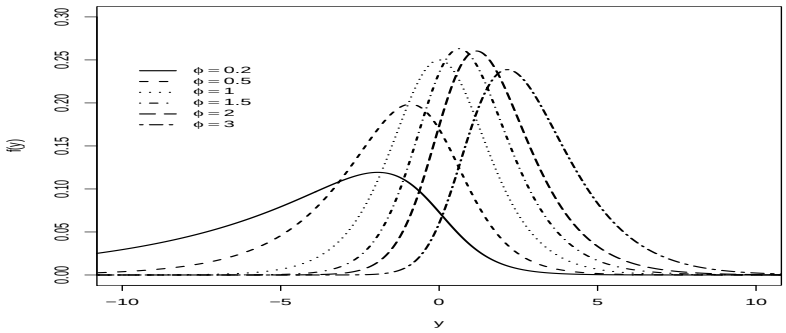

Fig. 2. Some shapes of the pdf of the LGL distribution.

$$\pi(z;\phi) = \frac{1}{\Gamma(\phi)} \exp(z) [1+\exp(z)]^{-2} \{\log[1+\exp(z)]\}^{\phi-1}, \quad -\infty < z < \infty. \tag{2.4}$$

The case $\phi = 1$ refers to the logistic distribution.

The rth moment of the LGL distribution is given by

$$\mu'_r = \frac{1}{\Gamma(\phi)} \sum_{i=0}^{r} \binom{r}{i} \sigma^i \mu^{r-i} \int_0^1 \left[\log\left(\frac{1-u}{u}\right)\right]^i \left[\log\left(\frac{1}{u}\right)\right]^{\phi-1} du, \tag{2.5}$$

where $u = \left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-1}$.

From equation (2.5) it is possible to determine the mean and variance of the LGL distribution.

2.1. *Characterizations of LGL Distribution*

We emphasize the importance of characterizing the new model LGL. Characterizations of the distributions provide conditions under which the underlying distribution is indeed that particular distribution of interest.

2.1.1. *Characterizations based on truncated moments*

In this subsection we present characterizations of LGL distribution in terms of a simple relationship between two truncated moments.

Proposition 2.1. *Let $Y : \Omega \to \mathbb{R}$ be a continuous random variable and let*

$$h(y) = g(y)\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-1} \quad \text{and} \quad g(y) = \left\{\log\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{1-\phi}, \quad \text{for} \quad y \in \mathbb{R}.$$

Then, pdf of Y is (3) if and only if the function η defined in Theorem A.1 (see Appendix) has the form

$$\eta(y) = 2\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right], \quad y \in \mathbb{R}. \tag{2.6}$$

Proof. Let Y have pdf (2.6), then

$$[1-F(y)]\,E[h(Y) \mid Y \geq y] = \frac{1}{2\Gamma(\phi)}\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-2}, \quad y \in \mathbb{R},$$

and

$$[1-F(y)]\,E[g(Y) \mid Y \geq y] = \frac{1}{\Gamma(\phi)}\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-1}, \quad y \in \mathbb{R},$$

and finally,

$$\eta(y)h(y) - g(y) = g(y) > 0 \text{ for } y \in \mathbb{R}.$$

Conversely, if $\eta(y)$ is given by (2.6), then

$$s'(y) = \frac{\eta'(y)h(y)}{\eta(y)h(y) - g(y)} = \frac{2}{\sigma}\exp\left(\frac{y-\mu}{\sigma}\right)\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-1},$$

from which we obtain

$$s(y) = 2\log\left\{\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}, \ y \in \mathbb{R}.$$

□

Now, in view of Theorem A.1 , Y has pdf (3).

Corollary 2.1. *Let $Y : \Omega \to \mathbb{R}$ be a continuous random variable and let h be as in Proposition 1.1. Then, pdf of Y is (3) if and only if there exist functions g and η defined in Theorem A.1 satisfying the following differential equation*

$$s'(y) = \frac{\eta'(y)h(y)}{\eta(y)h(y) - g(y)} = \frac{2}{\sigma}\exp\left(\frac{y-\mu}{\sigma}\right)\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-1}, \ y \in \mathbb{R}. \tag{2.7}$$

Remarks. (a) The general solution of the differential equation (2.7) is

$$\eta(y) = \left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^2\left[-\int\frac{2}{\sigma}\exp\left(\frac{y-\mu}{\sigma}\right)\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-3}(h(y))^{-1}g(y)\,dx + D\right],$$

for $y \in \mathbb{R}$, where D is a constant. One set of function (h,g,η) satisfying the above equation is given in Proposition 1.1 for $D = 0$.

(b) Clearly there are other triple of functions (h,g,η) satisfying the conditions of Theorem A.1. We presented one such pair in Proposition 1.1.

2.1.2. *Characterization based on hazard function*

The hazard function, ζ_F, of a twice differentiable distribution function, F, satisfies the first order differential equation

$$\frac{\zeta_F'(y)}{\zeta_F(y)} - \zeta_F(y) = q(y),$$

where $q(y)$ is an appropriate integrable function. Although this differential equation has an obvious form since

$$\frac{\zeta_F'(y)}{\zeta_F(y)} - \zeta_F(y) = \frac{f'(y)}{f(y)}, \tag{2.8}$$

for many univariate continuous distributions (2.8) seems to be the only differential equation in terms of the hazard function. The goal of the characterization based on hazard function is to establish a differential equation in terms of hazard function, which has as simple form as possible and is not of the trivial form (2.8). Here, we present a characterization of the of LGL model based on a nontrivial differential equation in terms of the hazard function.

Proposition 2.2. *Let $Y : \Omega \to \mathbb{R}$ be a continuous random variable. Then, Y has pdf* (3) *if and only if its hazard function ζ_F satisfies the differential equation*

$$\zeta_F'(y) - \frac{1}{\sigma}\zeta_F(y) = \exp\left(\frac{y-\mu}{\sigma}\right)\frac{d}{dx}\{B(y)\}, \quad y \in \mathbb{R} \tag{2.9}$$

where

$$B(y) = \frac{\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-2}\left\{\log\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\phi-1}}{\Gamma(\phi) - \gamma\left(\log\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right], \phi\right)}.$$

Proof. If Y has pdf (3), then clearly (2.9) holds. Now, if (2.9) holds, then after multiplying both sides of (2.9) by $\exp\left\{-\left(\frac{y-\mu}{\sigma}\right)\right\}$, we arrive at

$$\frac{d}{dy}\left\{\exp\left\{-\left(\frac{y-\mu}{\sigma}\right)\right\}\zeta_F(y)\right\} = \frac{d}{dy}\{B(y)\},$$

from which we have

$$\zeta_F(y) = \frac{f(y)}{1-F(y)} = \exp\left(\frac{y-\mu}{\sigma}\right)B(y), \quad y \in \mathbb{R}. \tag{2.10}$$

Integrating both sides of (2.10) from $-\infty$ to y, we have

$$F(x) = \frac{1}{\Gamma(\phi)}\gamma\left(\log\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right], \phi\right), \quad y \in \mathbb{R}.$$

□

3. The LGL regression models with censored data

In practice there are many situations where the response variable y_i is influenced by one or more explanatory variables or covariates, and can be related to treatments, intrinsic traits of sample units, exogenous variables, interactions or time-dependent variables, among others. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ be the explanatory variable vector associated with the ith response variable y_i for $i = 1, \ldots, n$.

For the first time, we construct a linear regression model for the response variable y_i based on the LGL distribution given by

$$y_i = \mathbf{x}_i^\top \beta + \sigma z_i,\ i = 1,\ldots,n, \tag{3.1}$$

where the random error z_i has the density function (2.4), $\beta = (\beta_1,\ldots,\beta_p)^\top$, $\sigma > 0$ and $\phi > 0$ are unknown scale parameters and $\mathbf{x}_i$ is the vector of explanatory variables modeling the location parameter $\mu_i = \mathbf{x}_i^\top \beta$. Hence, the location parameter vector $\mu = (\mu_1,\ldots,\mu_n)^\top$ of the LGL model has a linear structure $\mu = \mathbf{x}\beta$, where $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)^\top$ is a known model matrix. The logistic regression model is defined by (3.1) with $\phi = 1$.

Consider a sample $(y_1,\mathbf{x}_1),\ldots,(y_n,\mathbf{x}_n)$ of n independent observations, where each random response is defined by $y_i = \min\{\log(t_i),\log(c_i)\}$. We consider non-informative censoring such that the observed lifetimes and censoring times are independent. The log-likelihood function for θ reduces to

$$\begin{aligned} l(\theta) = -r\log(\sigma) - r\log[\Gamma(\phi)] + \frac{1}{\sigma}\sum_{i=1}^{n}\delta_i(y_i - \mathbf{x}_i^\top\beta) - 2\sum_{i=1}^{n}\delta_i\log\left[1+\right. \\ \left.\exp\left(\frac{y_i-\mathbf{x}_i^\top\beta}{\sigma}\right)\right] + (\phi-1)\sum_{i=1}^{n}\delta_i\log\left\{\log\left[1+\exp\left(\frac{y_i-\mathbf{x}_i^\top\beta}{\sigma}\right)\right]\right\} + \\ \sum_{i=1}^{n}(1-\delta_i)\log\left\{1-\frac{1}{\Gamma(\phi)}\gamma\left(\log\left[1+\exp\left(\frac{y_i-\mathbf{x}_i^\top\beta}{\sigma}\right)\right],\phi\right)\right\}, \end{aligned} \tag{3.2}$$

where $\theta = (\phi,\sigma,\beta^\top)^\top$ is the vector of unknown parameters, r is the number of non-censored observations.

The MLE $\widehat{\theta}$ of θ can be obtained by maximizing the log-likelihood function (3.2). We use the matrix programming language Ox (MaxBFGS function) (see Doornik, 2007) to calculate $\widehat{\theta}$. Initial values for β and σ are taken from the fit of the logistic regression model with $\phi = 1$.

From the fitted model (3.1), the survival function for y_i can be estimated by

$$\hat{S}(y;\hat{\phi},\hat{\sigma},\hat{\beta}^\top) = 1 - \frac{1}{\Gamma(\hat{\phi})}\gamma\left(\log\left[1+\exp\left(\frac{y-\mathbf{x}^\top\hat{\beta}}{\hat{\sigma}}\right)\right],\hat{\phi}\right).$$

Under general regularity conditions, the asymptotic distribution of $(\widehat{\theta}-\theta)$ is multivariate normal $N_{p+2}(0,K(\theta)^{-1})$, where $K(\theta)$ is the expected information matrix. The asymptotic covariance matrix $K(\theta)^{-1}$ of $\widehat{\theta}$ can be approximated by the inverse of the $(p+2)\times(p+2)$ observed information matrix $J(\theta)$ and then the inference on the parameter vector θ can be based on the normal approximation $N_{p+2}(0,J(\theta)^{-1})$ for $\widehat{\theta}$. The multivariate normal $N_{p+2}(0,J(\theta)^{-1})$ distribution can be used to construct approximate confidence regions for some parameters in θ and for the hazard and survival functions. In fact, a $100(1-\alpha^*)\%$ asymptotic confidence interval for each parameter θ_q is given by

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

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

likelihood ratio (LR) statistic can be used to discriminate between the LGL and logistic regression models since they are nested models. In this case, the hypotheses to be tested are $H_0 : \phi = 1$ versus $H_1 : H_0$ *is not true*, and the LR statistic reduces to $w = 2\{l(\widehat{\theta}) - l(\widetilde{\theta})\}$, where $\widetilde{\theta}$ is the MLE of θ under H_0. The null hypothesis is rejected if $w > \chi^2_{1-\alpha^*}(1)$, where $\chi^2_{1-\alpha^*}(1)$ is the quantile of the chi-square distribution with two degrees of freedom.

3.1. *Bootstrap re-sampling method*

The bootstrap re-sampling method was proposed by Efron (1979). The method considers the observed sample as if it represents the population. From the information obtained from the 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 mean, variance, percentiles and so on.

According to the literature, the re-sampling method may be non-parametric or parametric. In this study, the non-parametric bootstrap method is addressed, according to which the distribution function F can be estimated by the empirical distribution $\hat{F}$ (Hashimoto et al., 2013).

Let $\mathbf{T}=(T_1,\ldots,T_n)$ be an observed random sample and $\hat{F}$ be the empirical distribution of $\mathbf{T}$. Thus, a bootstrap sample $\mathbf{T}^*$ is constructed by re-sampling with replacement of n elements of the sample $\mathbf{T}$. For the B bootstrap samples generated, $T_1^*,\ldots,T_B^*$, the bootstrap replication of the parameter of interest for the bth sample is

$$\hat{\theta}_b^* = s(T_b^*),$$

i.e., the value of $\hat{\theta}$ for sample T_b^*, $b = 1,\ldots,B$.

The bootstrap estimator of the standard error (Efron and Tibshirani, 1993) is the standard deviation of these 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 $\bar{\theta}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$. Note that B is the number of bootstrap generated samples. According to Efron and Tibshirani (1993), assuming $B \geq 200$, it is generally sufficient to present good results to determine the bootstrap estimates. However, to achieve greater accuracy, a reasonably high B value must be considered. We describe the bias corrected and accelerated (BCa) method for constructing approximated confidence intervals based on the bootstrap re-sampling method.

4. Sensitivity analysis

4.1. *Global influence*

A first tool to perform sensitivity 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 the data set. A global influence measure considered by Xie and Wei (2007) is a generalization of Cook distance, which is defined as a standardized norm of $\hat{\theta}_{(i)} - \hat{\theta}$, and it is given by

$$GD_i(\theta) = (\hat{\theta}_{(i)} - \hat{\theta})^\top \left[\ddot{\mathbf{L}}(\theta)\right](\hat{\theta}_{(i)} - \hat{\theta}), \tag{4.1}$$

where $\ddot{\mathbf{L}}(\theta)$ is the observed information matrix.

Another measure to evaluate the influence of a case is presented by Cook and Weisberg (1982). This measure is called the likelihood distance and considers the difference between $\hat{\theta}_{(i)}$ and $\hat{\theta}$. Thus, the likelihood distance is given by

$$LD_i(\theta) = 2\Big[l(\hat{\theta}) - l(\hat{\theta}_{(i)})\Big], \tag{4.2}$$

where $l(\hat{\theta})$ is the value of the logarithm of the likelihood function for the full sample and $l(\hat{\theta}_{(i)})$ is the value of the logarithm of the likelihood function for the sample excluding the ith observation.

4.2. *Local influence*

This approach is suggested by Cook (1986), where instead of removing observations, weights are given to them. Local influence calculation can be carried out for model (3.1). If likelihood displacement $LD(\omega) = 2\{l(\hat{\theta}) - l(\hat{\theta}_{\omega})\}$ is used, where $\hat{\theta}_{\omega}$ denotes the MLE under the perturbed model, the normal curvature for θ at the direction $\mathbf{d}, \| \mathbf{d} \| = 1$, is given by $C_{\mathbf{d}}(\theta) = 2|\mathbf{d}^{\top}\Delta^{\top}\left[\ddot{\mathbf{L}}(\theta)\right]^{-1}\Delta\mathbf{d}|$, where Δ is a $(p+2)n$ matrix that depends on the perturbation scheme, whose elements are given by $\Delta_{vi} = \partial^2 l(\theta|\omega)/\partial\theta_v\partial\omega_i$, $i = 1,2,\ldots,n$ and $v = 1,2,\ldots,p+2$, evaluated at $\hat{\theta}$ and ω_0, where ω_0 is the no perturbation vector.

We can also calculate normal curvatures $C_{\mathbf{d}}(\phi)$, $C_{\mathbf{d}}(\sigma)$ and $C_{\mathbf{d}}(\beta)$ to perform various index plots, for instance, the index plot of $\mathbf{d}_{max}$, the eigenvector corresponding to $C_{\mathbf{d}_{max}}$, the largest eigenvalue of the matrix $\mathbf{B} = -\Delta^{\top}\left[\ddot{\mathbf{L}}(\theta)\right]^{-1}\Delta$ and the index plots of $C_{\mathbf{d}_i}(\phi)$, $C_{\mathbf{d}_i}(\sigma)$ and $C_{\mathbf{d}_i}(\beta)$, named total local influence (see, for example, Lesaffre and Verbeke, 1998), where $\mathbf{d}_i$ denotes an $n \times 1$ vector of zeros with one at the ith position. Thus, the curvature at direction $\mathbf{d}_i$ takes the form $C_i = 2|\Delta_i^{\top}\left[\ddot{\mathbf{L}}(\theta)\right]^{-1}\Delta_i|$, where $\Delta_i^{\top}$ denotes the ith row of Δ. It is customary to point out those cases such that $C_i \geq 2\bar{C}$, where $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$.

Next, we calculate, for three perturbation schemes (case-weight perturbation, response perturbation, explanatory variable perturbation), the matrix

$$\Delta = \Big(\Delta_{vi}\Big)_{\left[(p+2)\times n\right]} = \left(\frac{\partial^2 l(\theta|\omega)}{\partial\theta_v\partial\omega_i}\right)_{\left[(p+2)\times n\right]},$$

where $v = 1,2,\ldots,p+2$ and $i = 1,2,\ldots,n$, considering the model defined in (3.1) and its log-likelihood function given by (3.2).

5. Residual analysis

In this section, we compare two residuals to assess departures from the error assumptions and to detect outliers in the LGL regression model with censored observations. In the literature, various residuals, were investigated, for example, Collett (2003), Weisberg (2005) and Colosimo and Giolo (2006). In the context of survival analysis, the deviance residuals have been more widely used because they take into account the information of censored times (Silva et al., 2011). These residuals can also be used for the log-logistic regression model, which is a special case of the LGL regression model. Thus, the plot of the deviance residuals versus the observed times provides a way to verify the adequacy of the fitted model and to detect atypical observations.

The deviance residual can be expressed as

$$r_{D_i} = \begin{cases} \text{sign}(\hat{r}_{M_i})\Big\{-2\Big[1+\log\Big\{1-\frac{1}{\Gamma(\hat{\phi})}\gamma\Big[\log[1+\exp(\hat{z}_i)],\hat{\phi}\Big]\Big\}+ \\ \log\Big\{-\log\Big\{1-\frac{1}{\Gamma(\hat{\phi})}\gamma\Big[\log[1+\exp(\hat{z}_i)],\hat{\phi}\Big]\Big\}\Big\}\Big]\Big\} & \text{if} \quad \delta_i = 1, \\ \\ \text{sign}(\hat{r}_{M_i})\Big\{-2\log\Big\{1-\frac{1}{\Gamma(\hat{\phi})}\gamma\Big[\log[1+\exp(\hat{z}_i)],\hat{\phi}\Big]\Big\}\Big\}^{1/2} & \text{if} \quad \delta_i = 0, \end{cases} \tag{5.1}$$

where $\hat{r}_{M_i} = \delta_i + \log[\hat{S}(y_i;\widehat{\theta})]$ is the martingale residual, sign() is a function that leads to the values $+1$ if the argument is positive and -1 if the argument is negative and $\hat{z}_i = (y_i - \mathbf{x}_i^\top \hat{\beta})/\hat{\sigma}$.

5.1. *Simulation study*

A simulation study was carried out to investigate the behavior of the empirical distribution of the martingale residuals and the deviance components, and also to assess the MLEs in the LGL regression model with censored data.

For the simulation study, the variables $z_1,\ldots,z_n$ of the LGL distribution are generated by the acceptance-rejection method. More details can be found in Ross (2006). Therefore, for the sample sizes $n = 100$, $n = 300$ and $n = 500$, the values of the parameters of the distributions are set at $\phi = 0.8$ and 1.5, $\sigma = 1.0$, $\beta_0 = 2.0$ and $\beta_1 = 4.0$.

The survival times are generated from the following algorithm adapted for censored data:

1. Generate $v \sim$ uniform (a_1, b_1).
2. Generate $u \sim$ uniform $(0, b_2)$.
3. If $u \leq f(v) \Rightarrow z = v$, where $f(.)$ is the pdf (2.4).
4. Generate $x_1 \sim$ uniform $(0,1)$.
5. Write $y^* = \beta_0 + \beta_1 x_1 + \sigma z$.
6. Generate $c \sim$ uniform $(0,\tau)$, where τ was adjusted to obtain the percentages of right censoring 10% and 30%.
7. Write $y = \min(y^*, c)$ and $\delta = I(y^* \leq c)$ as the indicator of censoring.
8. Otherwise, return to step 1.

Therefore, 1,000 samples are generated for each combination of n, ϕ, σ and censoring percentage by using Monte Carlo simulations, and the MLEs of the model parameters are obtained for each of the samples. Then, for each fitted model, the residuals r_{D_i} in equation (5.1) are calculated. On the other hand, Figure 3 display the plots of the residuals versus the expected values of the order statistics of the standard normal distribution. This type of plot is known as the normal probability plot and serves to assess the departure from the assumption of normality of the residuals (Weisberg, 2005). Therefore, the following interpretations are obtained from these plots:

- The empirical distribution of the deviance residual agrees with the standard normal distribution.
- The empirical distribution of the martingale residual in general has accentuated asymmetry.
- When the censoring percentage decreases, the empirical distribution of the deviance residual appears to approach the standard normal distribution more rapidly than does the martingale residual.
- As the sample size increases, the empirical distribution of the deviance residual becomes closer to the normal distribution.

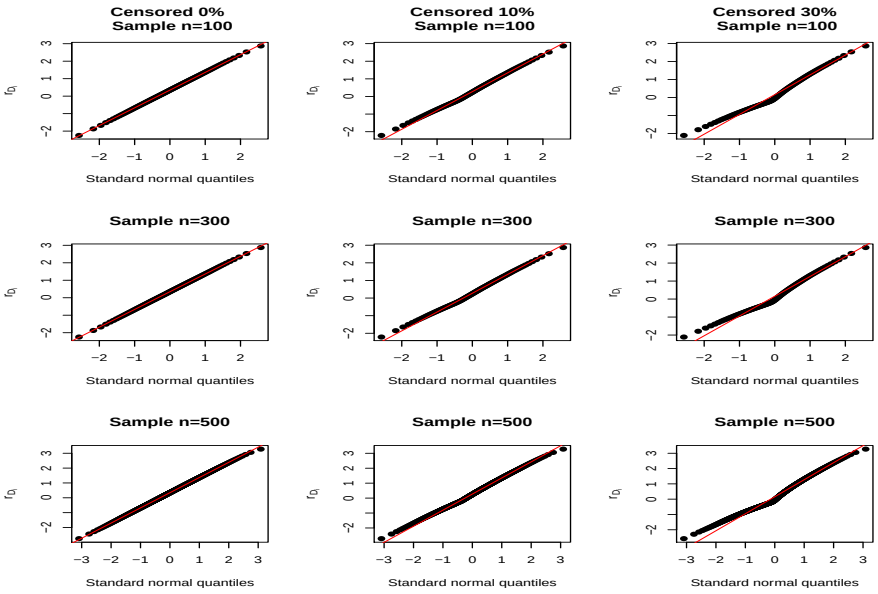

Fig. 3. QQ-plot for the deviance residual (r_{D_i}) of the LGL regression model with $\phi = 0.8$.

6. Application

To illustrate the application of the LGL regression model, we use a data set of Lee and Wang (2003). The data set refers to a study of patients diagnosed with acute myeloid leukemia (AML), where the objective was to determine factors associated with the lifetime of these patients. Thus, it was observed that 30 patient lifetimes and possible prognostic factors:

- x_1: Age (0= age of the patient $<$ 50 years, 1= age of the patient $\geq$ 50 years);
- x_2: Cellularity (0= cellularity of marrow clot section is 100%, 1=otherwise).

First, to verify the behavior of the AML data, we construct the Kaplan-Meier and the TTT curve (Aarset, 1987) displayed in Figure 4. From these plots, we note that the TTT-plot indicates that the lifetime of patients has a unimodal failure rate (Figure 4a), which justifies the LL distribution.

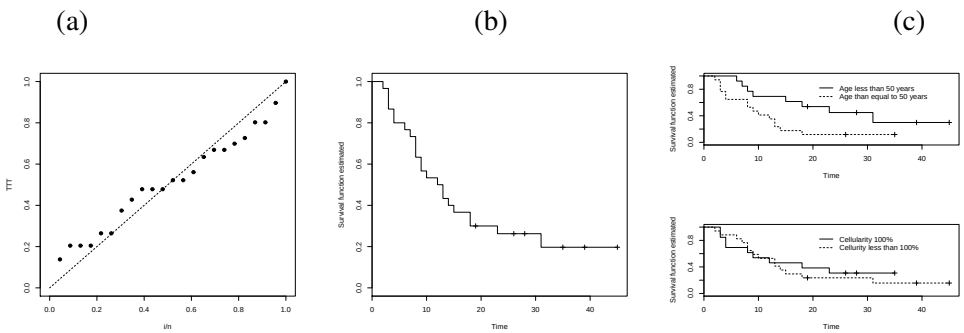

Fig. 4. (a) TTT-plot. Survival curve estimate by Kaplan-Meier method for the: (b) lifetime. (c) explanatory variable (age and cellularity).

Table 1. MLEs and non-parametric bootstrap estimates for the parameters of the regression model fitted to the AML data.

	MLEs			Non-parametric bootstrap		
Parameter	Estimate	S.E.	p-value	Estimate	S.E.	95% C.I. BCa
ϕ	5.918	4.529	–	5.069	0.910	(4.591, 7.940)
σ	0.397	0.172	–	0.389	0.041	(0.333, 0.465)
β_0	1.053	0.992	0.288	1.195	0.388	(0.431, 1.597)
β_1	-1.092	0.338	0.001	-1.071	0.330	(-1.582, -0.494)
β_2	-0.315	0.338	0.351	-0.315	0.336	(-0.810, 0.283)

Moreover, the time does not present a level above zero, as shown in Figure 4b. On the other hand, we have evidence that only the age of the patients have an influence on lifetime (Figure 4c). Thus, accordance whit is observed in Figure 4, we consider the following model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma z_i, \quad i = 1, \ldots, 30,$$

where $y_i = \log(t_i)$ denotes the logarithm of lifetime.

To maximize the log-likelihood function (3.2) and thus obtain the MLEs of the parameters of the proposed model, we use the subroutine `MaxBGFS` of the matrix programming language `Ox` version (`7.00`) with initial values $\phi = 1.0$, $\sigma = 0.760$, $\beta_0 = 4.880$, $\beta_1 =$ -0.153 and $\beta_2 = 0.026$. These values are obtained from the fit of the logistic regression model with censored data using the `R` software version (`2.15.1`).

Thus, in Table 1 we provide the parameter estimates, standard errors and significance of the MLEs and non-parametric bootstrap estimates. The figures in this table indicate that the estimates by the two methods are very similar, and then there is evidence that the presence of the only covariate x_1 (age) is significant at 5% level of significance.

Further, we calculate the maximum unrestricted and restricted log-likelihoods and the likelihood ratio (LR) statistics for testing a special model. For example, the LR statistic for testing the hypotheses H_0: $\phi = 1$ versus H_1: H_0 *is not true*, i.e., to compare the LGL and logistic regression models, is $w = 2(-35.548 + 37.540) = 3.924$ (p-value=0.048), which yields favorable indications toward to the LGL regression model.

The next step is to detect possible influential points in the LGL regression model. The measurements of global and local influence are calculated using the matrix programming language `Ox` version `7.00`.

The generalized Cook's distance (4.1) and likelihood distance (4.2) are given in Figure 5. These plots indicate that the points ♯8, ♯17, ♯18 and ♯30 are possible influential observations.

As for the plots of local influence, considering perturbations of cases ($C_{\mathbf{d}_{max}} = 1.1291$) and the logarithm of lifetime perturbation ($C_{\mathbf{d}_{max}} = 0.7453$), it is noted that the cases ♯5, ♯7, ♯8, ♯17, ♯18 and ♯30 can be considered as possible influential observations, as illustrated in Figure 6.

Also, Figure 7a indicates the deviance residuals. It can be noted that these residuals are randomized around zero, indicating the suitability of the model for analyzing the lifespan data of patients with AML.

Therefore, the sensitivity analysis (global influence and local influence) and the residual analysis detected that the influential observations are ♯17, ♯18 and ♯30, which appeared more frequently.

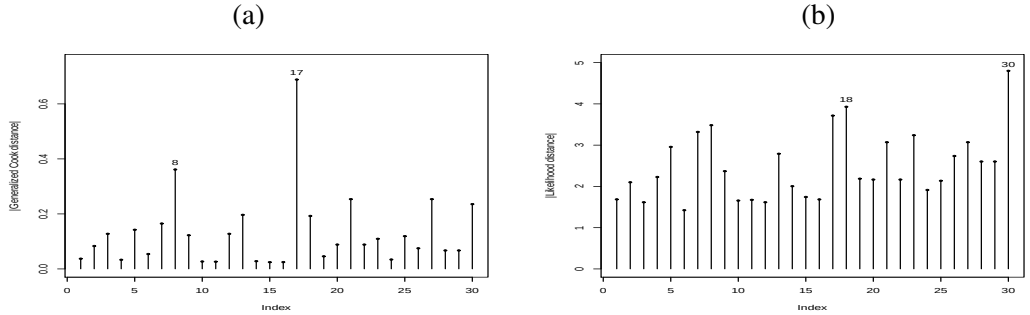

Fig. 5. Index plot of global influence from the LGL regression model fitted to the AML data. (a) Generalized Cook distance. (b) Likelihood distance.

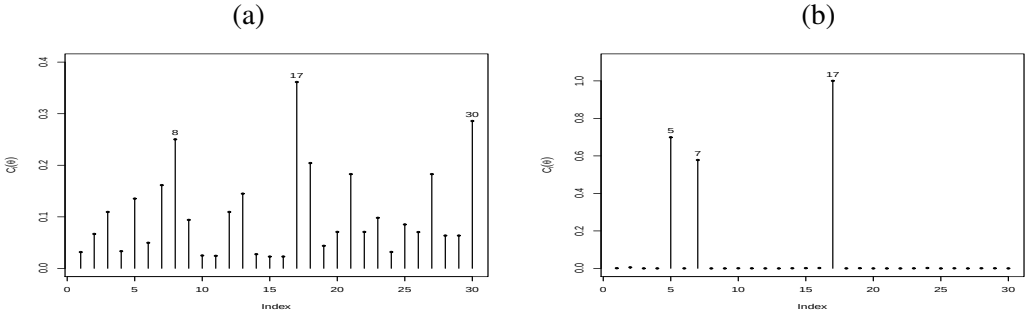

Fig. 6. Index plot of total local influence C_i from the LGL regression model fitted to the AML data. (a) Case-weight perturbation. (b) Response variable perturbation.

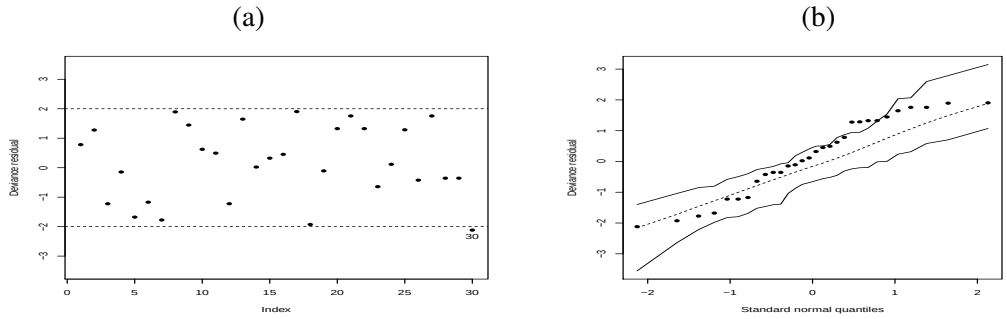

Fig. 7. (a) Deviance residual analysis for the LGL regression model fitted to the AML data. (b) Normal probability plot for the deviance residuals with envelopes.

These observations are identified as potential influential points, and correspond to patients who have the following descriptions:

- The observation ♯17 corresponds to the patient in the group of individuals with more than 50 years with the lowest lifetime.
- The observation ♯18 corresponds to the patient in the group of individuals with more than 50 years remained alive until the time 26.
- The observation ♯30 corresponds to the patient in the group of individuals with more than 50 who remained alive until the time 35.

Table 2. Relative changes [-**RC**-in %], estimates and corresponding *p*-values in parentheses for the regression coefficients to explain the log-survival times.

Set	$\hat{\phi}$	$\hat{\sigma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
A	-	-	-	-	-
	5.918	0.397	1.053	-1.092	-0.315
	(-)	(-)	(0.288)	(0.001)	(0.351)
A-{♯17}	[-2]	[4]	[15]	[18]	[63]
	6.063	0.383	0.899	-0.894	-0.117
	(-)	(-)	(0.376)	(0.012)	(0.741)
A-{♯18}	[-5]	[8]	[1]	[-2]	[19]
	6.190	0.367	1.040	-1.116	-0.255
	(-)	(-)	(0.303)	(0.001)	(0.438)
A-{♯30}	[-10]	[12]	[5]	[-2]	[21]
	6.529	0.351	1.001	-1.116	-0.249
	(-)	(-)	(0.342)	(0.001)	(0.444)
A-{♯17, ♯18}	[-9]	[12]	[17]	[16]	[81]
	6.460	0.349	0.871	-0.913	-0.058
	(-)	(-)	(0.399)	(0.007)	(0.865)
A-{♯17, ♯30}	[-17]	[17]	[23]	[16]	[83]
	6.946	0.330	0.815	-0.914	-0.053
	(-)	(-)	(0.457)	(0.006)	(0.875)
A-{♯18, ♯30}	[98]	[70]	[-258]	[-14]	[206]
	0.139	0.120	3.769	-1.243	0.334
	(-)	(-)	(0.000)	(0.000)	(0.077)
A-{♯17, ♯18, ♯30}	[70]	[-254]	[-11]	[51]	[211]
	0.146	0.117	3.731	-1.213	0.351
	(-)	(-)	(0.000)	(0.000)	(0.057)

Thus, to analyze the impact of these observations on the parameter estimates, we fit the model by eliminating individually each observation, and then removing two observations. Thus, in Table 2, we present the relative changes (in percentages) of each estimated parameter defined by $\mathbf{RC}_{\theta_j} = \left[(\hat{\theta}_j - \hat{\theta}_{j(i)})/\hat{\theta}_j\right] \times 100$, where $\theta_{j(i)}$is the MLE without the ith observation.

In view of the results of Table 2, it is noted that the MLEs of the parameters of the LGL regression model are not robust to the deletion of influential observations. Moreover, the significance of estimated parameters change (at the 5% significance level) after removal of the cases, that is, no changes inferential after removal of observations considered influential in diagnostics plots.

Finally, we verify the quality of the adjustment range of the LGL regression model by constructing the normal probability plot for the component of the waste diversion with simulated envelope (Atkinson, 1985). From this plot, we note that there is evidence of a good fit of the LGL regression model, as illustrated in Figure 7b.

Finally, the model fitted to the data is given by

$$\hat{y}_i = 0.719 - 0.972 x_{1i}, \quad i = 1, \ldots, 30. \tag{6.1}$$

Based on the final model (6.1), one can note that the variability in the lifetime of patients with AML can be explained by means of their age. In other words, patients aged 50 and older have a 38% lower chance of surviving than patients younger than 50 years. However, for sensitivity analysis, it is necessary to have two patients older than 50 who remained alive until the end of the study. In this

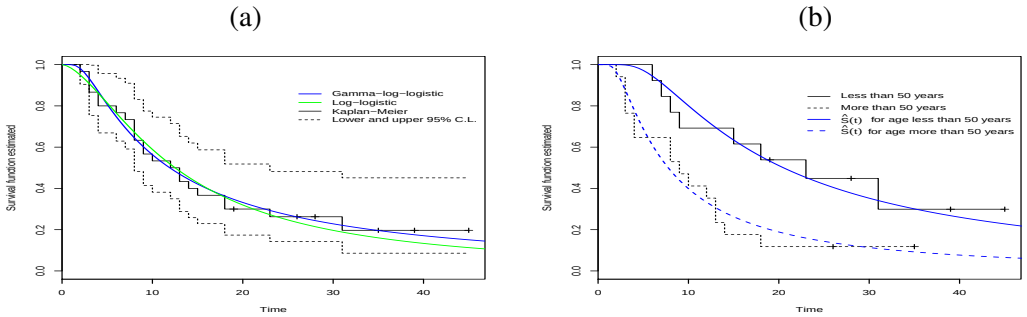

Fig. 8. (a) Estimated survival function by fitting the GLL distribution and the empirical survival for the AML data. (b) Estimated survival function for the LGL regression model and the empirical survival.

case, it would be interesting to carry out another study to verify the causes for the different behavior of these patients in relation to other patients with AML.

Goodness of fit. In order to assess if the model is appropriate, the plot comparing the empirical survival function and estimated survival function for the GLL distribution is displayed in Figure 8a. A graphical comparison for the LGL regression model is given in Figure 8b. The curves displayed in this figure are the empirical survival and estimated survival functions determined from (8). Based on this figure, we note that both models show satisfactory fits. However, the LGL model presents a better fit to the current data.

Moreover, according to these plots, it can be noted that the LGL regression model can be used to predict the chances of survival of a patient diagnosed as suffering from AML. In other words, the chance that a patient younger than 50 will survive until the 15th period is approximately 63% $[\hat{S}(15|x_1 = 0) = 0.634]$, while the chance for a patient 50 or older is only 26% $[\hat{S}(15|x_1 = 1) = 0.263]$. It can also be verified that the chance that a patient aged 50 or older will survive to the 40h period is very small $[\hat{S}(40|x_1 = 1) = 0.077]$. Therefore, according to the patient age group, it is possible to determine the best form for treating patients with AML.

7. Concluding Remarks

A three parameter lifetime model called the gamma-log-logistic (GLL) distribution is proposed. It is a simple generalization of the log-logistic distribution. The new model extends several distributions widely used in the lifetime literature and it is more flexible than the Weibull, log-logistic and log-normal distributions. Based on this new distribution, we propose a log-gamma logistic (LGL) regression model, which is more suitable for modeling censored and uncensored lifetime data. The proposed model serves as an important extension to several existing regression models and could be a valuable addition to the literature. Hence, the proposed regression model is a good alternative for lifetime data analysis and can be more flexible than the log-logistic, log-normal and Weibull models. Several simulation studies are performed for different parameter settings, sample sizes and censoring percentages. Maximum likelihood and parametric bootstrapping are described for estimating the model parameters. Diagnostic analysis is presented to assess local and global influences. We also discuss the sensitivity of the maximum likelihood estimates (MLEs) from the fitted model using deviance component residuals and sensitivity analysis. The usefulness of the proposed regression model is also demonstrated by means of a real data set.

Acknowledgments

This work was supported by FAPESP grant 2010/04496-2, Brazil.

Appendix

Theorem A.1. *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a given probability space and let $H = [a,b]$ be an interval for some $a < b$ ($a = -\infty$, $b = \infty$might as well be allowed). Let $Y : \Omega \to H$ be a continuous random variable with the distribution function F and let h and g be two real functions defined on H such that*

$$\frac{E\left[g\left(Y\right) \mid Y \geq y\right]}{E\left[h\left(Y\right) \mid Y \geq y\right]} = \eta\left(y\right), \quad y \in H$$

is defined with some real function η. Assume that $h, g \in C^1(H)$, $\eta \in C^2(H)$ and F is twice continuously differentiable and strictly monotone function on the set H. Finally, assume that the equation $\eta h = g$ has no real solution in the interior of H. Then F is uniquely determined by the functions h, g and η, particularly

$$F\left(y\right) = \int_a^y C \left| \frac{\eta'\left(u\right)}{\eta\left(u\right)h\left(u\right) - g\left(u\right)} \right| \exp\left(-s\left(u\right)\right)\, du\,,$$

where the function s is a solution of the differential equation $s' = \frac{\eta' h}{\eta h - g}$ and C is a constant, chosen to make $\int_H dF = 1$.

References

[1] C. Alexander, G.M. Cordeiro, E.M.M. Ortega and J.M. Sarabia, Generalized beta-generated distribution, *Computational Statistics and Data Analysis* **56** (2012) 1880–1897.
[2] A.C. Atkinson, *Plots, transformations and regression: an introduction to graphical methods of diagnostic regression analysis*, (University Press, Oxford, 1985).
[3] M.V. Aarset, How to identify bathtub hazard rate, *IEEE Transactions on Reliability* **36** (1987) 106–108.
[4] D. Collett, *Modelling survival data in medical research*, (Chapman & Hall, Boca Raton, London, New York, 2003).
[5] E.A. Colosimo and S.R. Giolo, *Anlise de sobrevivłncia aplicada*, (Edgard Blcher, So Paulo, 2006).
[6] R.D. Cook, Assessment of local influence (with discussion), *Journal of the Royal Statistical Society B* **48** (1986) 133–169.
[7] R.D. Cook and S. Weisberg, *Residuals and influence in regression*, (Chapman & Hall, Boca Raton, London, New York, 1982).
[8] G.M. Cordeiro and M. de Castro, A new family of generalized distributions, *Journal of Statistical Computation and Simulation* **81** (2011) 883–898.
[9] J.A. Doornik, *An Object-Oriented Matrix Language Ox 5*, (Timberlake Consultants Press, London, 2007).
[10] B. Efron, Bootstrap methods: another look at the jackknife, *The Annals of Statistics* **7** (1979) 1–26.
[11] B. Efron and R. Tibshirani, *An introduction to the bootstrap*, (Chapman & Hall, Boca Raton, London, New York, 1993).
[12] N. Eugene, C. Lee and F. Famoye, Beta-normal distribution and its applications, *Communications in Statistics - Theory and Methods* **31** (2002) 497–512.
[13] E.M. Hashimoto, G.M. Cordeiro and E.M.M. Ortega, The new Neyman type A beta Weibull model with long-term survivors, *Computational Statistics* **28** (2013) 933–954.
[14] J.F. Lawless, *Statistical models and methods for lifetime data*. (John Wiley & Sons, New York, 2003).
[15] E.T. Lee and J.W. Wang, *Statistical methods for survival data analysis*, 3rd Edition, (John Wiley & Sons, New Jersey, 2003).

[16] M.W.A. Ramos, G.M. Cordeiro, P.R.D. Marinho, C.R.B. Dias and G.G. Hamedani, The Zografos-Balakrishnan log-logistic distribution: properties and applications, *Journal of Statistical Theory and Applications* **12** (2013) 225–244.
[17] M.M. Ristic and N. Balakrishnan, The Gamma Exponentiated Exponential Distribution, *Journal of Statistical Computation and Simulation* **82** (2012) 1191–1206.
[18] S.M. Ross, *Simulation*, (Elsevier Academic Press, Boston, 2006).
[19] T.V.F. Santana, E.M.M. Ortega, G.M. Cordeiro and G.O. Silva, The Kumaraswamy-log-logistic distribution, *Journal of Statistical Theory and Applications* **11** (2012) 265–291.
[20] Z. Shayan, S.M.T. Ayatollahi and N. Zare, A parametric method for cumulative incidence modeling with a new four-parameter log-logistic distribution, *Theoretical Biology and Medical Modeling* **8** (2011) 1–11.
[21] M.M. Shoukri, I.U.M. Mian and D.S. Tracy, Sampling properties of estimators of the log-logistic distribution with application to Canadian precipitation data, *The Canadian Journal of Statistics* **16** (1988) 223–236.
[22] G.O. Silva, E.M.M. Ortega and G.A. Paula, Residuals for log-Burr XII regression models in survival analysis, *Journal of Applied Statistics* **38** (2011) 1435–1445.
[23] S. Weisberg, *Applied linear regression*, (John Wiley & Sons, New York, 2005).
[24] F.C. Xie and B.C. Wei, Diagnostics analysis in censored generalized Poisson regression model, *Journal of Statistical Computation and Simulation* **77** (2007) 695–708.
[25] N. Zare, M. Doostfatemeh and A. Rezaianzadeh, Modeling of breast cancer prognostic factors using a parametric log-logistic model in Fars Province, Southern Iran, *Asian Pacific Journal of Cancer Prevention* **13** (2012) 533–1537.
[26] K. Zografos and N. Balakrishnan, On families of beta-and generalized gama-generated distributions and associated inference, *Statistical Methodology* **6** (2009) 344–362.