5-1-2016

New Flexible Regression Models Generated by Gamma Random Variables with Censored Data

Elizabeth M. Hashimoto
Universidade Federal do Parana

Gauss M. Cordeiro
Universidade Federal de Pernambuco-Brazil

Edwin M. M. Ortega
Universidade de Sao Paulo

Gholamhossein G. Hamedani
Marquette University, gholamhoss.hamedani@marquette.edu

Published version. *International Journal of Statistics and Probability,* Vol. 5, No. 3 (May 2016): 9-31. DOI.

International Journal of Statistics and Probability; Vol. 5, No. 3; May 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

New Flexible Regression Models Generated by Gamma Random Variables with Censored Data

Elizabeth M. Hashimoto[1], Gauss M. Cordeiro[2], Edwin M.M. Ortega[3] & G.G. Hamedani[4]

[1] Departamento Acadêmico de Matemática, UTFPR, Londrina, Brazil

[2] Departamento de Estatística, UFPE, Recife, Brazil

[3] Departamento Ciências Exatas, ESALQ-USP, Piracicaba, Brazil

[4] Departament of Mathematics, Statistics and Computer Science, Milwaukee, USA

Correspondence: Edwin M.M. Ortega, Departamento de Cilncias Exatas, USP, Av. Pdua Dias 11 - Caixa Postal 9, 13418-900, Piracicaba-S£o Paulo, Brazil. E-mail: edwin@usp.br

Received: November 3, 2016 Accepted: February 4, 2016 Online Published: April 8, 2016

doi:10.5539/ijsp.v5n3p9 URL: http://dx.doi.org/10.5539/ijsp.v5n3p9

Abstract

We propose and study a new log-gamma Weibull regression model. We obtain explicit expressions for the raw and incomplete moments, quantile and generating functions and mean deviations of the log-gamma Weibull distribution. We demonstrate that the new regression model can be applied to censored data since it represents a parametric family of models which includes as sub-models several widely-known regression models and therefore can be used more effectively in the analysis of survival data. We obtain the 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 distribution 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 a modified deviance residual 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-Weibull distribution, regression model, residual analysis, sensitivity analysis

1. Introduction

The CAPES *(Coordenação de Aperfeiçoamento de Pessoal de Nível Superior)* is responsible for assessing every three years, postgraduate courses in the country with the goal of keeping the courses with a level of excellence and contribute to the training of researchers (Horta and Moraes, 2005) in Brazil. According to CAPES, one of the requirements is that the master programs are done in two years and doctoral courses in four years (Moreira et al., 2010). For this reason, the programs graduate has studied how to encourage students before or within the stipulated time. To determine the best strategy to take, a particular program postgraduate noted the time it took his doctoral students to hold, according to sex and age at first registration.

One way to study the effect of these explanatory variables (gender and age of the students) on the completion time is through a location-scale regression model, also known as a model of accelerated lifetime. These models consider that the response variable belongs to a family of distributions characterized by a parameter location and scale parameter. Further details on this class of regression models can be found in Cox and Oakes (1984), Kalbfleisch and Prentice (2002) and Lawless (2003).

However, for the case of parametric models, it is assumed that the time until the end of the PhD is sampled from a continuous distribution. In the context of survival analysis, some distributions have been used to analyze censored data. For example, Leiva et al. (2007) conducted a diagnostic study in a log-Birnbaum-Saunders regression model. Carrasco et al. (2008) defined a regression model considering a modification of the Weibull distribution. Ortega et al. (2011) proposed the beta-Weibull regression model and Silva et al. (2011) proposed the log-Burr XII regression model. Thus, using the same approach adopted in this work, a distribution obtained from a generated gamma family will be expressed in the form of models belonging to the location-scale models. In this way, we can study the influence of explanatory variables on the completion time of doctoral students.

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. In this paper, we propose a location-scale regression model based on the *log-gamma Weibull* (LGW) distribution. We consider a classic analysis for the LGW regression model. The inferential part was carried out using asymptotic distribution of the maximum likelihood estimators (MLEs), which, in situations when the sample is small, may present difficult results to be justified. As an alternative to classic analysis we explore the use of bootstrap method for survival times analysis as a feasible alternative. After modeling, it is important to check assumptions in the model and to conduct a robustness study in order to detect influential or extreme observations that can cause distortions in the results of the analysis. Numerous approaches have been proposed in the literature to detect influential or outlying observations. On the other hand, when using case deletion, all information from a single subject is deleted at once and, therefore, it is hard to tell whether that subject has any influence on a specific aspect of the model. A solution for the earlier problem can be found in a quite different paradigm, being a local influence approach where one again investigates how the results of an analysis are changed under small perturbations in the model, and where these perturbations can be specific interpretations. We develop a similar methodology to detect influential subjects in LGW regression models with censored data. Further, we compare two residuals to assess departures from the error assumptions and to detect outlying observations in the LGW 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.

In Section 2, we perform a brief review on the GW and LGW distributions and derive the quantile function (qf) of the last distribution. Some of the mathematical properties of the LGW model such as the ordinary and incomplete moments, mean deviations, Bonferroni and Lorenz curves and generating function are investigated in Section 3. In Section 4, we consider a brief study on the GW distribution and present certain the characterizations of LGW distribution. In Section 5, we obtain the MLEs and the estimates based on a bootstrap method and provide some results from simulation studies for the LGW regression model with censored data. In Section 6, we use diagnostic measures considering three perturbation schemes and case-deletion in the LGW regression model with censored observations. Section 7 deals with the definition and discussion of the residuals and presents useful comments on the results from various simulation studies. In Section 8, a real data set is analyzed for illustrative purposes. Finally, we offer some conclusions in Section 9.

2. The Log-gamma-Weibull Distribution

The art of proposing generalized distributions has attracted theoretical and applied statisticians due to their flexible properties. Most of the distributions used to describe real data are chosen for the following reasons: a physical or statistical theoretical argument to explain the mechanism of the generated data, a model that has previously been used successfully and an appropriate model whose empirical fit is good to data. The Weibull distribution is a very popular distribution for modeling lifetime data and for modeling phenomenon with monotone failure rates. When modeling monotone hazard rates, the Weibull distribution may be an initial choice because of its negatively and positively skewed density shapes. This distribution has cumulative distribution function (cdf) (for $t > 0$) given by

$$G(t;\alpha,\lambda) = 1 - \exp\Big[-\Big(\frac{t}{\lambda}\Big)^{\alpha}\Big], \tag{1}$$

where $\alpha > 0$ is a shape parameter and $\lambda > 0$ a scale parameter. The probability density function (pdf) corresponding to (1) is given by

$$g(t;\alpha,\lambda) = \frac{\alpha}{\lambda^{\alpha}}\, t^{\alpha-1} \exp\Big[-\Big(\frac{t}{\lambda}\Big)^{\alpha}\Big]. \tag{2}$$

We write $T \sim \mathrm{GW}(\alpha,\lambda)$ for a random variable T having the pdf (2).

There has been an increased interest in defining new univariate continuous distributions by introducing one additional shape parameter to the baseline distribution. The extra parameter has been proved useful in some cases to explore tail properties and to improve the goodness-of-fit of the new distribution. In fact, Zografos and Balakrishnan (2009) and Ristic, and Balakrishnan (2012) proposed a family of univariate distributions generated by one-parameter gamma random variables. Based on any baseline cdf $G(t)$, they defined the *gamma-G family* with pdf $f(t)$ and cdf $F(t)$ (for $\phi > 0$) given by

$$f(t) = \frac{1}{\Gamma(\phi)}\,\{-\log[1-G(t)]\}^{\phi-1}\; g(t) \tag{3}$$

and

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

respectively, where $g(t) = dG(t)/dt$, $\Gamma(a) = \int_0^{\infty} u^{a-1}\,\mathrm{e}^{-\mathrm{u}}\mathrm{du}$ is the gamma function, and $\gamma(a,z) = \int_0^{z} u^{a-1}\,\mathrm{e}^{-\mathrm{u}}\mathrm{du}$ is the incomplete gamma function. The gamma-G family has the same parameters of the G distribution plus one additional

shape parameter $\phi > 0$. For any specified G distribution, we can generate the associated gamma-G distribution. Recently Nadarajah *et al.* (2015) provide a comprehensive treatment of general mathematical properties of gamma-G distributions. 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).

In this context, we define the *gamma Weibull* (GW) density function by inserting (1) and (2) in equation (3). So, we obtain

$$f(t) = \frac{\alpha}{\lambda^{\phi\,\alpha}\,\Gamma(\phi)}\, t^{\phi\alpha-1}\, \exp\Big[-\Big(\frac{t}{\lambda}\Big)^{\alpha}\Big], \quad t > 0, \tag{4}$$

where $\phi > 0$ and $\alpha > 0$ are shape parameters and $\lambda > 0$ is a scale parameter. By combining the gamma-G and Weibull distributions, we obtain the generalized gamma distribution (Stacy, 1962). The applications of the GW distribution can be directed to model insurance data, tree diameters, software reliability, extreme value observations in floods, carbon fibrous composites, firmware system failure, reliability prediction and fracture toughness, among others.

To obtain a distribution that belongs to the location-scale model, we consider the transformation of the random variable $Y = \log(T)$, which has the *log-gamma Weibull* (LGW) distribution. Thus, considering the pdf (4) and the transformations $\alpha = \sigma^{-1}$ and $\lambda = \mathrm{e}^{\mu}$, the LGW distribution (for $y \in \mathrm{R}$) reduces to

$$f(y) = f(y;\phi,\mu,\sigma) = \frac{1}{\sigma\Gamma(\phi)}\, \exp\left[\frac{\phi(y-\mu)}{\sigma} - \exp\Big(\frac{y-\mu}{\sigma}\Big)\right], \tag{5}$$

where $\phi > 0$, $\sigma > 0$ and $\mu \in \mathrm{R}$. Therefore, Y follows the LGW distribution. Plots of the pdf (5) for selected parameter values are displayed in Figure 1. These plots show great flexibility for different values of the shape parameter ϕ with $\mu = 0$ and $\sigma = 1$. If Y is a random variable having pdf (5), we write $Y \sim \mathrm{LGW}(\phi,\mu,\sigma)$.

If $T \sim \mathrm{GW}(\phi,\alpha,\lambda)$, then $Y = \log(T) \sim \mathrm{LGW}(\phi,\mu,\sigma)$. The survival function corresponding to (5) becomes

$$S(y) = S(y;\phi,\mu,\sigma) = 1 - \frac{1}{\Gamma(\phi)}\gamma\Big[\exp\Big(\frac{y-\mu}{\sigma}\Big),\phi\Big]. \tag{6}$$

The simulation of Y is very easy: if V is a gamma random variable with shape parameter ϕ and unit scale parameter then $Y = \mu + \sigma\,\log(V)$, will have the LGW density function (5).

Let $z = Q^{-1}(a,u)$ be the inverse function of $Q(a,z) = 1 - \gamma(a,x)/\Gamma(a) = u$, see http:// functions.wolfram.com/ GammaBetaErf/ InverseGammaRegularized/ for details. The asymptotes of $z = Q^{-1}(a,u)$ can be determined using known properties of $Q^{-1}(a,u)$. Using http:// functions.wolfram.com/ GammaBetaErf/ InverseGammaRegularized/ 06/ 02/ 01/, we can write $z = Q^{-1}(a,u)$ as one can see that

$$z = Q^{-1}(a,u) = \left[-(1-a)\,W_{-1}\left(-\frac{(1-u)^{1/(a-1)}\Gamma(a)^{1/(a-1)}}{a-1}\right)\right]$$

as $u \to 0$, where $W_{-1}(\cdot)$ denotes the product log function. Further, using http:// functions.wolfram.com/ GammaBetaErf/InverseGammaRegularized/06/01/03/0001/, we have

$$z = Q^{-1}(a,u) = \left[-(1-u)^{1/a}\,\Gamma(a+1)^{1/a} + \frac{(1-u)^2\,\Gamma(a+1)^{2/a}}{(a+1)}\right] + O((1-u)^3).$$

Further, inverting $1 - S(y) = u$, we obtain the qf of Y as

$$y = \mu + \sigma\,\log\Big[Q^{-1}(a,u)\Big]. \tag{7}$$

We define the standardized random variable $Z = (Y-\mu)/\sigma$ having pdf given by

$$\pi(z;\phi) = \frac{1}{\Gamma(\phi)}\exp\{\phi\,z - \exp(z)\}, \qquad z \in \mathrm{R}. \tag{8}$$

The special case $\phi = 1$ corresponds to the log-Weibull (LW) (or extreme-value) distribution and, for $\sigma = 1$, we obtain the log-gamma-exponential (LGE) model.

3. Mathematical Properties

In this section, we derive explicit expressions for the ordinary and incomplete moments, generating function and Bonferroni and Lorenz curves for the LGW distribution. Its mathematical properties are not difficult to be implemented in applications because of the computational and analytical facilities available in programming softwares like MATHEMATICA and MAPLE that can easily tackle the problems involved in computing the special functions in these properties.

3.1 Moments

The need for necessity and the importance of moments in any statistical analysis especially in applied work is obvious. Some of the most important features and characteristics of a distribution can be studied through moments (e.g. tendency, dispersion, skewness and kurtosis). The rth raw moment of Y can be expressed as

$$\mu'_r = E(Y^r) = \frac{1}{\Gamma(\phi)} \sum_{i=0}^{r} \binom{r}{i} \sigma^i \mu^{r-i} \int_0^\infty [\log(u)]^i \, u^{\phi-1} \, \mathrm{e}^{-\mathrm{u}} \mathrm{du}, \tag{9}$$

where $u = \exp\left(\frac{y-\mu}{\sigma}\right)$. Further, based on a result by Prudnikov et al. (1986, equation 2.6.21.1), we can rewrite (9) as

$$\mu'_r = \frac{1}{\Gamma(\phi)} \sum_{i=0}^{r} \binom{r}{i} \sigma^i \mu^{r-i} \frac{\partial^i \Gamma(\phi)}{\partial \phi^i}. \tag{10}$$

The mean and variance of Y follow from (10) as

$$E(Y) = \mu'_1 = \mu + \sigma\psi(\phi)$$

and

$$\mathrm{Var}(Y) = E^2(X) = \sigma^2\psi(1,\phi) + \sigma^2\psi^2(\phi),$$

where $\psi(\cdot)$ is the digamma function, $\psi(n,\phi)$ is the polygamma function and n is a positive integer.

The central moments (μ_r) and cumulants (κ_r) of Y can be determined from (10) as

$$\mu_r = \sum_{k=0}^{r} (-1)^k \binom{r}{k} \mu_1'^k \mu'_{r-k} \qquad \text{and} \qquad \kappa_r = \mu'_r - \sum_{k=1}^{r-1} \binom{r-1}{k-1} \kappa_k \mu'_{r-k},$$

respectively, where $\kappa_1 = \mu'_1$. Then, $\kappa_2 = \mu'_2 - \mu'^2_1$, $\kappa_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1$, $\kappa_4 = \mu'_4 - 4\mu'_3\mu'_1 - 3\mu'^2_2 + 12\mu'_2\mu'^2_1 - 6\mu'^4_1$, etc. The skewness $\gamma_1 = \kappa_3/\kappa_2^{3/2}$ and kurtosis $\gamma_2 = \kappa_4/\kappa_2^2$ follow from the second, third and fourth cumulants. Other kinds of moments such L-moments may also be obtained in closed-form, but we consider only the previous moments for reasons of space.

Figure 2 provides some plots of the skewness (Figure 2a) and kurtosis (Figure 2b) for $\sigma = 1.5$ as a function of μ for some values of ϕ, whereas Figure 3 provides some plots of the skewness (Figure 3a) and kurtosis (Figure 3b) for $\mu = 0$ as a function of ϕ for some values of σ. It can be noted that when the parameter μ increases the LGW distribution is higher and concentrated than the standard normal distribution, i.e., it has heavy tails. Moreover, when the parameter ϕ decreases, the LGW distribution becomes positive asymmetrical (Figure 2). On the other hand, Figure 3 indicates that when the parameter ϕ increases, the shape of the LGW distribution is flatter than that of the normal distribution. For the parameter ϕ greater than three, this distribution has a negative skewness and if the parameter ϕ is less than three, the distribution is positively skewed.

The nth descending factorial moment of Y is

$$\mu'_{(n)} = E(Y^{(r)}) = E\left[Y(Y-1) \times \cdots \times (Y-r+1)\right] = \sum_{k=0}^{r} s(r,k)\, \mu'_k,$$

where

$$s(r,k) = (k!)^{-1} \left[\frac{d^k \, x^{(r)}}{dx^k} \right]_{x=0}$$

is the Stirling number of the first kind which counts the number of ways to permute a list of r items into k cycles. So, we can obtain the factorial moments from the ordinary moments given before.

3.2 Mean Deviations

For empirical purposes, the shapes of many distributions can be usefully described by what we call the first incomplete moment, which plays an important role for measuring inequality, for example, income quantiles and Lorenz and Bonferroni curves. The first incomplete moment of Y is given by $m_1(z) = \int_{-\infty}^{z} y\, f(y;\phi,\mu,\sigma)dy$. Changing variable $u = \exp\left(\frac{y-\mu}{\sigma}\right)$ and using a similar approach of Section 3.1, we can write

$$m_1(z) = \frac{1}{\Gamma(\phi)} \left[\mu\, \gamma(\phi, z^\star) + \sigma\, J(\phi, z^\star)\right], \tag{11}$$

where $z^\star = \exp\left(\frac{z-\mu}{\sigma}\right)$, $\gamma(\phi, z^\star) = \int_0^{z^\star} u^{\phi-1}\, \mathrm{e}^{-\mathrm{u}}\mathrm{du}$ is the incomplete gamma function, and $J(\phi, z^\star) = \int_0^{z^\star} \log(u)\, u^{\phi-1}\, \mathrm{e}^{-\mathrm{u}}\mathrm{du}$.

The integral $J(\phi, z^\star)$ can be expressed in terms of the digamma function $\psi(\phi) = d\log[\Gamma(\phi)]/d\phi$ and the Meijer G-function defined by

$$G_{p,q}^{m,n}\left(x \,\middle|\, \begin{array}{c} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{array}\right) = \frac{1}{2\pi \mathrm{i}} \int_L \frac{\prod_{j=1}^{m} \Gamma\left(b_j + t\right) \prod_{j=1}^{n} \Gamma\left(1 - a_j - t\right)}{\prod_{j=n+1}^{p} \Gamma\left(a_j + t\right) \prod_{j=m+1}^{q} \Gamma\left(1 - b_j - t\right)}\, x^{-t}\, dt,$$

where $\mathrm{i} = \sqrt{-1}$ is the complex unit and L denotes an integration path (see Gradshteyn and Ryzhik, 2000; Section 9.3) for a description of this path. The Meijer G-function contains many integrals with elementary and special functions. Some of these integrals are included in Prudnikov et al. (1986). In the MATHEMATICA software, $\psi(\phi)$ is denoted by PolyGamma$[0, \phi]$ and the Meijer G-function is represented by

$$G_{p,q}^{m,n}\left(x \,\middle|\, \begin{array}{c} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{array}\right) = \text{MeijerG}\left[\{\{\mathrm{a}_1, \ldots, \mathrm{a}_\mathrm{n}\}, \{\mathrm{a}_{\mathrm{n}+1}, \ldots, \mathrm{a}_\mathrm{p}\}\}, \{\{\mathrm{b}_1, \ldots, \mathrm{b}_\mathrm{m}\}, \{\mathrm{b}_{\mathrm{m}+1}, \ldots, \mathrm{b}_\mathrm{q}\}\}, \mathrm{x}\right].$$

We can obtain using the MATHEMATICA software

$$J(\phi, z^\star) = -\gamma(\phi, z^\star)\, \log(z^\star) - \text{MeijerG}\left[\{\{\}, \{1, 1\}\}, \{\{0, 0, \phi\}, \{\}\}, \mathrm{z}^\star\right] + \Gamma(\phi)\, \text{PolyGamma}[0, \phi].$$

Applications of equation (11) can be addressed to obtain Bonferroni and Lorenz curves defined for a given probability π by $B(\pi) = m_1(q)/(\pi\mu_1')$ and $L(\pi) = m_1(q)/\mu_1'$, respectively, where $\mu_1' = E(Y)$ and $q = Q(\pi)$ is the qf (7) of Y at π.

The mean deviations about the mean ($\delta_1 = E(|Y - \mu_1'|)$) and about the median ($\delta_2 = E(|Y - M|)$) of Y can be expressed as

$$\delta_1 = 2\mu_1'\, F(\mu_1') - 2m_1(\mu_1') \qquad \text{and} \qquad \delta_2 = \mu_1' - 2m_1(M),$$

respectively, where $\mu_1' = E(Y)$, $M = Median(Y) = Q(0.5)$ is the median computed from (7), $F(\mu_1') = 1 - S(\mu_1')$ is easily calculated from the survival function (6) and $m_1(z)$ is given by (11).

3.3 Generating Function

The moment generating function (mgf) provides the basis of an alternative route to analytical results compared with working directly with the pdf and cdf and it is widely used in the characterization of distributions and the application of the skew-normal test (Meintanis, 2010) and other goodness of fit tests (Ghosh, 2013). Therefore, using a result in Prudnikov et al. (1986, equation 2.6.21.1), we can derive the mgf of Y as

$$M_Y(t) = \frac{\mathrm{e}^{\mathrm{t}\mu}}{\Gamma(\phi)} \sum_{i=0}^{\infty} \frac{(t\sigma)^i}{i!}\, \frac{\partial^i}{\partial\phi^i}\Gamma(\phi).$$

4. Characterizations of LGW Distribution

In this section we present certain characterizations of LGW distribution. The first characterization is based on a simple relationship between two truncated moments. It should be mentioned that for this characterization, the cdf need no have a closed form. We believe, due to the nature of the cdf of LGW, there may not be other possibly interesting characterizations than the ones presented in this section. Our first characterization result borrows from a theorem due to [*Glä*zel, 1987] , see Theorem G in the Appendix A. Note that the result holds also when the interval I is not closed. Moreover, as shown in [*Glä*zel, 1990], this characterization is stable in the sense of weak convergence. Here is our first characterization of LGW distribution.

Proposition 1. Let $X : \Omega \to \mathbb{R}$ be a continuous random variable and let $q_1(x) = \exp\left\{(1-\phi)\left(\frac{x-\mu}{\sigma}\right)\right\}$ and $q_2(x) = q_1(x) \exp\left\{-\mathrm{e}^{\left(\frac{x-\mu}{\sigma}\right)}\right\}$ for $x \in \mathbb{R}$. The random variable X belongs to LGW family (5) if and only if the function η defined in Theorem G has the form

$$\eta(x) = \frac{1}{2} \exp\left\{-\mathrm{e}^{\left(\frac{x-\mu}{\sigma}\right)}\right\}, \quad x \in \mathbb{R}.$$

Proof. Let X be a random variable with density (5), then

$$(1 - F(x)) E[q_1(X) \mid X \geq x] = \frac{1}{\Gamma(\phi)} \exp\left\{-e^{\left(\frac{x-\mu}{\sigma}\right)}\right\}, \quad x \in \mathbb{R},$$

and

$$(1 - F(x)) E[q_2(X) \mid X \geq x] = \frac{1}{2\Gamma(\phi)} \exp\left\{-2e^{\left(\frac{x-\mu}{\sigma}\right)}\right\}, \quad x \in \mathbb{R},$$

and finally

$$\eta(x) q_1(x) - q_2(x) = -\frac{1}{2} q_1(x) \left\{-\exp\left\{-e^{\left(\frac{x-\mu}{\sigma}\right)}\right\}\right\} < 0 \quad for \ \ x \in \mathbb{R}.$$

Conversely, if η is given as above, then

$$s'(x) = \frac{\eta'(x) q_1(x)}{\eta(x) q_1(x) - q_2(x)} = \frac{1}{\sigma} e^{\left(\frac{x-\mu}{\sigma}\right)}, \quad x \in \mathbb{R},$$

and hence

$$s(x) = e^{\left(\frac{x-\mu}{\sigma}\right)}, \quad x \in \mathbb{R}.$$

Now, in view of Theorem G, X has density (5) .

Corollary 1. Let $X : \Omega \to \mathbb{R}$ be a continuous random variable and let $q_1(x)$ be as in Proposition 1. The pdf of X is (5) if and only if there exist functions q_2 and η defined in Theorem G satisfying the differential equation

$$\frac{\eta'(x) q_1(x)}{\eta(x) q_1(x) - q_2(x)} = \frac{1}{\sigma} e^{\left(\frac{x-\mu}{\sigma}\right)}, \quad x \in \mathbb{R}.$$

The general solution of the differential equation in Corollary 1 is

$$\eta(x) = \exp\left\{e^{\left(\frac{x-\mu}{\sigma}\right)}\right\} \left[\int \frac{1}{\sigma} \exp\left\{\left(\frac{x-\mu}{\sigma}\right) - e^{\left(\frac{x-\mu}{\sigma}\right)}\right\} (q_1(x))^{-1} q_2(x) \, dx + D\right],$$

where D is a constant. Note that a set of functions satisfying the differential equation in Corollary 1, is given in Proposition 1 with $D = 0$. However, it should be also noted that there are other triplets (q_1, q_2, η) satisfying the conditions of Theorem G.

5. The Log-gamma Weibull Regression Model with Censored Data

The last decade is full of works on generalized classes of regression models, which are always precious for applied statisticians. In practical applications, the lifetimes are affected by explanatory variables such as the cholesterol level, blood pressure and many 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 $i = 1, \ldots, n$.

We construct a linear regression model for the response variable y_i based on the LGW distribution given by

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \sigma z_i, \ \ i = 1, \ldots, n, \tag{12}$$

where the random error z_i has the pdf (8), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\sigma > 0$ and $\phi > 0$ are unknown scalar parameters and $\mathbf{x}_i$ is the vector of explanatory variables modeling the location parameter $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$. Hence, the location parameter vector

$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ of the LGW regression model has a linear structure $\boldsymbol{\mu} = \mathbf{x}\boldsymbol{\beta}$, where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is a known model matrix.

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 assume non-informative censoring such that the observed lifetimes and censoring times are independent. Let F and C be the sets of individuals for which y_i is the log-lifetime or log-censoring, respectively. We consider non-informative censoring such that the observed lifetimes and censoring times are independent. The log-likelihood function for the vector of parameters $\boldsymbol{\theta} = (\phi, \sigma, \boldsymbol{\beta}^\top)^\top$ from model (12) has the form $l(\boldsymbol{\theta}) = \sum_{i\in F} l_i(\boldsymbol{\theta}) + \sum_{i\in C} l_i^{(c)}(\boldsymbol{\theta})$, where $l_i(\boldsymbol{\theta}) = \log[f(y_i)]$, $l_i^{(c)}(\boldsymbol{\theta}) = \log[S(y_i)]$, $f(y_i)$ is the pdf (5) and $S(y_i)$ is the survival function (6), for $i = 1, \ldots, n$. Therefore, the log-likelihood function for $\boldsymbol{\theta}$ reduces to

$$l(\boldsymbol{\theta}) = -r\log(\sigma) - r\log[\Gamma(\phi)] + \frac{\phi}{\sigma}\sum_{i\in F}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}) - \sum_{i\in F}\exp\left(\frac{y_i - \mathbf{x}_i^\top\boldsymbol{\beta}}{\sigma}\right) + \sum_{i\in C}\log\left\{1 - \frac{1}{\Gamma(\phi)}\gamma\left[\exp\left(\frac{y_i - \mathbf{x}_i^\top\boldsymbol{\beta}}{\sigma}\right), \phi\right]\right\}, \tag{13}$$

where $\boldsymbol{\theta} = (\phi, \sigma, \boldsymbol{\beta}^\top)^\top$ is the vector of unknown parameters, r is the observed number of failures and $\gamma(x, \phi)$ is the incomplete gamma function. The components of the score vector $U(\boldsymbol{\theta})$ are given by

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} = -r\psi(\phi) + \frac{1}{\sigma}\sum_{i\in F}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}) - \sum_{i\in C}\frac{V(\exp(z_i), \phi, 1) - \psi(\phi)\gamma[\exp(z_i), \phi]}{\Gamma(\phi) - \gamma[\exp(z_i), \phi]},$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \sigma} = -\frac{r}{\sigma} - \frac{\phi}{\sigma}\sum_{i\in F} z_i + \frac{1}{\sigma}\sum_{i\in F} z_i\exp(z_i) + \frac{1}{\sigma}\sum_{i\in C}\frac{z_i\exp[\phi z_i - \exp(z_i)]}{\Gamma(\phi) - \gamma[\exp(z_i), \phi]}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} = -\frac{\phi}{\sigma}\sum_{i\in F} x_{ij} + \frac{1}{\sigma}\sum_{i\in F} x_{ij}\exp(z_i) + \frac{1}{\sigma}\sum_{i\in C}\frac{x_{ij}\exp[\phi z_i - \exp(z_i)]}{\Gamma(\phi) - \gamma[\exp(z_i), \phi]},$$

where $z_i = \frac{y_i - \mathbf{x}_i^\top}{\sigma}$, $\psi(\cdot)$ is the digamma function, $V(x, \phi, 1) = \int_0^x u^{\phi-1}e^{-u}\log(u)du$ and $j = 0, 1, \ldots, p$.

The MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained by maximizing the log-likelihood function (13). We use the matrix programming language Ox (MaxBFGS function) (see Doornik, 2007) to calculate the estimate $\widehat{\boldsymbol{\theta}}$. Initial values for $\boldsymbol{\beta}$ and σ are taken from the fit of the LW regression model with $\phi = 1$.

Under general regularity conditions, the asymptotic distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is multivariate normal $N_{p+3}(0, K(\boldsymbol{\theta})^{-1})$, where $K(\boldsymbol{\theta})$ is the expected information matrix. The asymptotic covariance matrix $K(\boldsymbol{\theta})^{-1}$ of $\widehat{\boldsymbol{\theta}}$ can be approximated by the inverse of the $(p + 3) \times (p + 3)$ observed information matrix $J(\boldsymbol{\theta})$ and then the inference on the parameter vector $\boldsymbol{\theta}$ can be based on the normal approximation $N_{p+3}(0, J(\boldsymbol{\theta})^{-1})$ for $\widehat{\boldsymbol{\theta}}$. This multivariate normal $N_{p+3}(0, J(\boldsymbol{\theta})^{-1})$ distribution can be used to construct approximate confidence regions for some parameters in $\boldsymbol{\theta}$ and for the hazard and survival functions.

5.1 Bootstrap Re-sampling Method

The bootstrap is a computer-based method for assessing the accuracy of statistical estimates and tests. It was first proposed by Efron (1979). Treat the data as if they were the(true, unknown) population and draw samples (with replacement) from the data as if you were sampling from the population. Repeat the procedure a large number of times(say B) each time computing the quantity of interest. Then, use the B values of the quantity of interest to estimate its unknown distribution.

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, $\mathbf{T}_1^*, \ldots, \mathbf{T}_B^*$, the bootstrap replication of the parameter of interest for the bth sample is given by

$$\hat{\boldsymbol{\theta}}_b^* = s(\mathbf{T}_b^*),$$

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

The bootstrap estimator of the standard error (Efron and Tibshirani, 1993) is the standard deviation of these bootstrap samples, namely

$$\widehat{EP}_B = \left[\frac{1}{(B-1)} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \bar{\theta}_B \right)^2 \right]^{1/2},$$

where $\bar{\theta}_B = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$. Note that B is the number of bootstrap samples generated. 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 (Hashimoto et al., 2013).

6. Sensitivity Analysis

There are basically two approaches to detecting observations that seriously influence the results of a statistical analysis. One approach is the case-deletion approach, and the second approach is one in which the stability of the estimated outputs with respect to the model inputs is studied via various minor model perturbation schemes.

6.1 Global Influence

The assessment of robustness aspects of the parameter estimates in statistical models has been an important concern of various researchers in recent decades. The case deletion measures , which consists of studying the impact on the parameter estimates after dropping individual observations, is probably the most employed technique to detect influential observations (Cook and Weisberg, 1982)

A global influence measure considered by Xie and Wei (2007) is a generalization of the Cook distance defined as a standardized norm $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$. It is expressed as

$$GD_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^\top [\ddot{\mathbf{L}}(\boldsymbol{\theta})](\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}), \tag{14}$$

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

Another measures to evaluate the influence is called of likelihood distance and considers the difference between $\hat{\boldsymbol{\theta}}_{(i)}$ and $\hat{\boldsymbol{\theta}}$.

Thus, the likelihood distance is given by

$$LD_i(\boldsymbol{\theta}) = 2\Big[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)}) \Big], \tag{15}$$

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

6.2 Local Influence

A second tool for sensitivity analysis is known as local influence. The local influence measures are calculated after a perturbation scheme in the model or data is established. Thus different perturbation schemes can be used according to the purpose of the analysis. Therefore, considering the model defined in (12) and the logarithm of the likelihood function expressed in (13), the following perturbation schemes are used:

a. ***Case perturbation***
Let $0 \leq \omega_i \leq 1$ and $\boldsymbol{\omega}_0 = (1, \ldots, 1)^\top$ be the vector representing no perturbation. The logarithm of the log-likelihood disturbed for each model incorporating different weights for each element of the data is defined by

$$l(\boldsymbol{\theta}) = -\{r \log(\sigma) - r \log[\Gamma(\phi)]\} \sum_{i \in F} \omega_i + \frac{\phi}{\sigma} \sum_{i \in F} \omega_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) - \sum_{i \in F} \Big[\omega_i \times \exp\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma} \right) \Big] + \sum_{i \in C} \omega_i \log \left\{ 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\exp\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma} \right), \phi \right] \right\}.$$

b. ***Response variable perturbation***
To verify the sensitivity of the model (12), it is assumed that the response variable $y_i = \log(t_i)(i = 1, \ldots, n)$ is submitted to the additive perturbation scheme such that $y_i^* = y_i + \omega_i \text{sd}(y_i)$, where $\text{sd}(y_i)$ is a scaling factor which can be the

standard deviation of the logarithm of failure times and $\omega_i \in \mathbf{R}$ is a perturbation vector (Silva et al., 2010). In this case, $\boldsymbol{\omega}_0 = (0, \ldots, 0)^\top$ is the vector representing no perturbation and the logarithm likelihood function disturbed is expressed as

$$l(\boldsymbol{\theta}) = -r\log(\sigma) - r\log\left[\Gamma(\phi)\right] + \frac{\phi}{\sigma}\sum_{i\in F}(y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta}) - \sum_{i\in F}\exp\left(\frac{y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta}}{\sigma}\right) + \sum_{i\in C}\log\left\{1 - \frac{1}{\Gamma(\phi)}\gamma\left[\exp\left(\frac{y_i^* - \mathbf{x}_i^\top\boldsymbol{\beta}}{\sigma}\right), \phi\right]\right\}.$$

c. ***Explanatory variable perturbation***
Another way of evaluating the sensitivity of the model (12) is to consider small perturbations in a particular continuous explanatory variable, denoted by X_j. In this case, the explanatory variable is submitted to the additive perturbation scheme, such that $x_{ij}^* = x_{ij} + \omega_i \mathrm{sd}(x_{ij})$, where$\mathrm{sd}(x_{ij})$ is scaling factor that can be the standard deviation of the disturbed explanatory variable (Silva et al., 2010). Thus, considering that $\mathbf{x}_i^{*\top}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_j x_{ij}^* + \ldots + \beta_p x_{ip}$ and $\boldsymbol{\omega}_0 = (0, \ldots, 0)^\top$ is the vector representing no perturbation, the logarithm of the likelihood function is given by

$$l(\boldsymbol{\theta}) = -r\log(\sigma) - r\log\left[\Gamma(\phi)\right] + \frac{\phi}{\sigma}\sum_{i\in F}(y_i - \mathbf{x}_i^{*\top}\boldsymbol{\beta}) - \sum_{i\in F}\exp\left(\frac{y_i - \mathbf{x}_i^{*\top}\boldsymbol{\beta}}{\sigma}\right) + \sum_{i\in C}\log\left\{1 - \frac{1}{\Gamma(\phi)}\gamma\left[\exp\left(\frac{y_i - \mathbf{x}_i^{*\top}\boldsymbol{\beta}}{\sigma}\right), \phi\right]\right\}.$$

For all three perturbation schemes, the array of maximum curvature is calculated numerically as

$$\boldsymbol{\Delta} = (\boldsymbol{\Delta}_{vi})_{(p+3)\times n} = \left[\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\theta_v \partial\boldsymbol{\omega}_i}\right]_{(p+3)\times n},$$

where $v = 1, \ldots, p+3$, $i = 1, \ldots, n$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ is the vector of weights that penalizes the LGW regression model or the observations.

7. Residual Analysis

In order to study the assumptions of the errors and the presence of outliers, we propose various residuals, for example, Collett (2003), Weisberg (2005) and Colosimo and Giolo (2006). In the context of survival analysis, the deviance residual has been more widely used, because they take into account the information of censored times (Silva et al., 2008). Thus, the deviance residual plot versus the observed times provides a way to verify the adequacy of the adjusted model and help to find atypical observations. The deviance residual is expressed as

$$r_{D_i} = \begin{cases} \mathrm{sign}(\hat{r}_{M_i})\Big\{-2\Big[1 + \log\Big\{1 - \frac{1}{\Gamma(\hat{\phi})}\gamma\Big[\log[1+\exp(\hat{z}_i)], \hat{\phi}\Big]\Big\} + \\ \log\Big\{-\log\Big\{1 - \frac{1}{\Gamma(\hat{\phi})}\gamma\Big[\log[1+\exp(\hat{z}_i)], \hat{\phi}\Big]\Big\}\Big\}\Big]\Big\} & \text{if} \quad i \in F, \\ \\ \mathrm{sign}(\hat{r}_{M_i})\left\{-2\log\left\{1 - \frac{1}{\Gamma(\hat{\phi})}\gamma\Big[\log[1+\exp(\hat{z}_i)], \hat{\phi}\Big]\right\}\right\}^{1/2} & \text{if} \quad i \in C, \end{cases} \tag{16}$$

where

$$r_{M_i} = \begin{cases} 1 + \log[\hat{S}(y_i; \widehat{\boldsymbol{\theta}})] & \text{if} \quad i \in F, \\ \\ \log[\hat{S}(y_i; \widehat{\boldsymbol{\theta}})] & \text{if} \quad i \in C, \end{cases}$$

is the martingale residuals, $\mathrm{sign}(\cdot)$ is a function that leads the values $+1$ if the argument is positive and -1 if the argument is negative and $\hat{z}_i = (y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}})/\hat{\sigma}$ for $i = 1, \ldots, n$.

7.1 Simulation Study

We performed a simulation to assess the MLEs of the LGW regression model with censored data, and also to investigate the behavior of the empirical distribution of martingale and deviance residuals. For the simulation study, the variables

$z_1, \ldots, z_n$ of the LGW distribution (8) were generated by the acceptance-rejection method, as explained in Ross (2006) and Bonat et al. (2012). 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 (Zeviani, 2012), adapted in this work for censored data.

i. Generate $v \sim \text{uniform}(a_1, b_1)$, where a_1 and b_1 are chosen to represent the support of random variable with pdf (8).

ii. Generate $u \sim \text{uniform}(0, b_2)$, where b_2 was chosen to represent the values of the density (8).

iii. If $u \leq f(v) \Rightarrow z = v$, where $f(\cdot)$ is the function (8).

iv. Generate $x_1 \sim \text{uniform}(0, 1)$.

v. Write $y^* = \beta_0 + \beta_1 x_1 + \sigma z$.

vi. Generate $c \sim \text{uniform}(0, \tau)$, where τ was adjusted to obtain the percentages of right censoring 10% and 30%.

vii. If $y = \min(y^*, c)$ then $y \in F(\text{Failure})$, else $y \in C(\text{Censored})$.

viii. Otherwise, return to step i.

Therefore, 1,000 samples are generated for each combination of n, ϕ, σ and censoring percentages by means of Monte Carlo simulations, and the MLEs of the model parameters are obtained for each of the samples. Then, for each adjusted model, the residuals r_{D_i} (16) are determined. On the other hand, Figures 4-5 display the plots of the residuals versus the expected values of the order statistics of the standard normal distribution. This plot is known as the normal probability plot and serves to assess the departure from the normality assumption 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 and as the sample size increases, the empirical distribution of the deviance residual becomes closer to the normal distribution (as illustrated in Figures 4-5).

8. Application

In this study, information from 49 PhD students of a particular program of postgraduate between the periods 1999 to 2012 are used. The interest of the study is to check whether the median completion time is within the maximum period (four years) stipulated by CAPES. Further, we wish to verify whether gender and age in the year of first registration exert some influence on the completion time of the students and if there is interaction between these explanatory variables. Thus, the response variable was defined as the time (in months) of first registration until the end of the PhD (date of defense). However, students who dropped out or who abandoned the course and did not return or who have not completed the course during this period are considered censored times. So, we define the following variables, y: logarithm of time (months), x_1: age (years) and x_2: gender (0= female, 1=male).

In many applications there is qualitative information about the hazard shape, which can help with selecting a particular model. In this context, a device called the total time on test (TTT) plot (Aarset 1987) is useful. The TTT plot is obtained by plotting $\mathsf{G}(r/n) = [(\sum_{i=1}^{r} T_{i:n}) + (n-r)T_{r:n}]/(\sum_{i=1}^{n} T_{i:n})$, where $r = 1, \ldots, n$ and $T_{i:n}$, $i = 1, \ldots, n$ are the order statistics of the sample, against r/n (Mudholkar *et al*, 1996).

First, to verify the behavior of the PhD data, the Kaplan-Meier and the TTT curves are displayed in Figure 6. From these plots, it is noted that the TTT-plot indicates that the time until the end of the PhD of students presents an increasing failure rate (Figure 6a). Moreover, the time not present a level above zero as shown in Figure 6b. There will be evidence that the gender of the students did affect the time until the end of the PhD. Thus, in accordance with what was observed in Figure 6, we can consider the following model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \sigma z_i, \quad i = 1, \ldots, 49.$$

To maximize the function (13) and obtain the MLEs of the parameters of the proposed model, we use the `MaxBGFS` subroutine of the matrix programming language `Ox` version `6.20` with initial values $\phi = 1.000$, $\sigma = 0.760$, $\beta_0 = 4.880$, $\beta_1 = -0.153$ and $\beta_2 = 0.026$, obtained from the fit of the Weibull regression model in software `R` (version `2.15.1`). Thus, Table 1 provides the parameter estimates of the models, standard errors and significance of the parameters for the MLEs and non-parametric bootstrap estimates. By examining the figures in this table, we conclude that the estimates by the two methods are very similar. So, there is an evidence that the presence of the interaction between age and gender at the time of completion of a PhD degree.

The next step is to detect possible influential points in the LGW regression model. The measurements of global and local influence are calculated using the matrix programming language `Ox` (version `6.20`). Generalized Cook's distance (14) and likelihood distance (15) are displayed in Figure 7. From this figure, it is noted that the cases ♯8, ♯18 and ♯21 are possible influential observations. For the local influence plots, considering perturbation of cases ($C_{\mathbf{d}_{max}} = 2.1805$), the logarithm of time perturbation ($C_{\mathbf{d}_{max}} = 0.3925$) and explanatory variable perturbation x_1 ($C_{\mathbf{d}_{max}} = 4.4124$), it is noted that the points ♯18 and ♯21 can be considered as possible influential observations as illustrated in Figures 8-9. On the other hand, the plot of the deviance residuals versus the fitted values is displayed in Figure 10. It is observed that there is evidence that the observations ♯7, ♯18 and ♯21 are discrepant (Figure 10a).

Therefore, the sensitivity analysis (global influence and local influence) and residual analysis detect points ♯18 and ♯21 as possible influential observations. These observations identified as potential influential points correspond to the students who have the following descriptions:

i. The observation 18 corresponds to a 31 years old female student, who defended her PhD in a maximum time (60 months).

ii. The observation 21 corresponds to a 31 years old male student, who defended his PhD in a minimum time (35 months).

Thus, to analyze the impact of these observations on the parameter estimates, we adjust the model eliminating individually each observation, and then removing the two observations. In Table 2, we present relative changes (in percentages) of the estimates 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.

On the figures of Table 2, we note that the MLEs of the parameters of the LGW 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 the cases, that is, no changes inferential after removal of observations considered influential in diagnostics plots. Therefore, the observations are kept in the data set.

Finally, we verify the quality of the adjustment range of the LGW regression model by constructing in Figure 11 the normal probability plot for the component of the waste diversion with simulated envelope (Atkinson, 1985). This figure reveals that there is evidence of a good fit of the LGW regression model to the current data. Thus, the fitted model to the data can be expressed as

$$\hat{y}_i = 4.1351 - 0.0045\, x_{i1} - 0,4237\, x_{i2} + 0,0101\, x_{i1}\, x_{i2}, \quad i = 1, \ldots, 49.$$

Thus, according to this model, the time until the end of the PhD for male students varies with age, in this case, older students have the highest degree of time than the younger students.

9. Concluding Remarks

In this paper, we derive explicit expressions for the raw and incomplete moments, quantile and generating functions and mean deviations of the log-gamma Weibull distribution. Based on this distribution, we construct a new log-gamma Weibull regression model to investigate the effect of the age and gender at the time until the end of a PhD student. We also provide diagnostic measures to test the adequacy of the fitted model. In particular, the estimates of the parameters of the new model obtained by two methods are similar. Further, we provide diagnostic analysis for the fitted model to the the times of the students until the end of their PhD. Considering these aspects, we conclude that the fitted model can explain the effect of the variables on the time and that there is an interaction between age and gender of the students at the time until the end of the PhD.

Acknowledgment

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

Appendix A

Theorem G. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a given probability space and let $I = [d, e]$ be an interval for some $d < e$ ($d = -\infty$, $e = \infty$ might as well be
Let $X : \Omega \to I$ be a continuous random variable with the distribution function F and let q_1 and q_2 be two real functions defined on I such that

$$\mathbf{E}\left[q_2(X) \mid X \geq x\right] = \mathbf{E}\left[q_1(X) \mid X \geq x\right] \eta(x), \quad x \in I,$$

is defined with some real function η. Assume that $q_1, q_2 \in C^1(H)$, $\eta \in C^2(H)$ and F is twice continuously differentiable and strictly monotone function on the set I. Finally, assume that the equation $q_1\eta = q_2$ has no real solution in the interior of I. Then F is uniquely determined by the functions q_1 , q_2 and η, particularly

$$F(x) = \int_d^x C \left| \frac{\eta'(u)}{\eta(u) q_1(u) - q_2(u)} \right| \exp(-s(u)) \, du ,$$

where the function s is a solution of the differential equation $s' = \frac{\eta' q_1}{\eta q_1 - q_2}$ and C is the normalization constant, such that $\int_I dF = 1$.

Appendix B

The Ox 6.21 (object-oriented matrix programming language) code for the simulation study.

B.1. LGLL model

```
#include<oxstd.h>
#include<oxdraw.h>
#include<oxprob.h>
#include<oxfloat.h>
#include<maximize.h>
#include<simula.h>
#pragma link("maximize.oxo")
static decl g_mT;
static decl n;
log_vero(const vP, const adFunc, const avScore, const amHessiam){
decl cont,y,x,xx1,xbeta,z;
decl uns=ones(n,1);
decl vero=zeros(1,n);
decl phi=vP[0];
decl sigma=vP[1];
decl beta0=vP[2];
decl beta1=vP[3];
//
for(cont=0;cont<n;++cont){
y=g_mT[cont][0];
xx1=g_mT[cont][2];
xbeta=beta0+(beta1*xx1);
z=(y-xbeta)/sigma;
x=log(1+exp(z));
//
if(g_mT[cont][1]==1){
vero[0][cont]=-log(sigma)-loggamma(phi)+z-(2*x)+((phi-1)*log(x+1e-10));
}
else{
vero[0][cont]=log(1-probgamma(x,phi,1)+1e-10);
}
}
adFunc[0]=double(vero*uns);
return 1;
}
main(){
ranseed("GM");

// Variaveis locais
decl i,j,theta0,g_mC,g_mZ,g_mY,g_mD,x1,pcens,cont,mu,xbeta,z,x,Sy;

// Numero de simulacoes
```

```
decl r=1000;

// Tamanho da mostra
n=100;

// Porcentagem de censura
pcens=0;

// Parametros da distribuicao
//decl phi1=0.8;
decl phi1=1.5;
decl sig1=1;
decl b0=2;
decl b1=4;
theta0=<1.5;1;2;4>;
//
decl theta=zeros(r,4,0);
decl rmi=zeros(n,r);
decl rm_ord=zeros(n,r);
decl rdi=zeros(n,r);
decl rd_ord=zeros(n,r);
decl nc=pcens*n;
decl testecens=zeros(n,r);
decl testey=zeros(n,r);
decl testex=zeros(n,r);

j=0;
do{
g_mC=zeros(n,1);
g_mZ=zeros(n,1);
g_mD=ones(n,2);
//
i=0;
while(i<n){
// Valores para phi=0.8
//decl xi=-10+(20*ranu(1,1));
//decl yi=0.237*ranu(1,1);
//
// Valores para phi=1.5
decl xi=-10+(20*ranu(1,1));
decl yi=0.265*ranu(1,1);
decl fdp=(1/gammafact(phi1))*exp(xi)*((1+exp(xi))^(-2))*((log(1+exp(xi)+1e-10))^(phi1-1));
if(yi<=fdp){
g_mZ[i]=xi;
i++;
}
}
x1=ranu(n,1);
mu=b0+(b1*x1);
g_mY=mu+(sig1*g_mZ);
//
cont=0;
for(i=0;i<n;i++){
if(cont<nc){
decl cte=5;
g_mC[i]=cte*ranu(1,1);
if(g_mY[i]<g_mC[i]){
```

```
g_mD[i][0]=g_mY[i];
}
else{
g_mD[i][0]=g_mC[i];
g_mD[i][1]=0;
cont=cont+1;
}
}
else{
g_mD[i][0]=g_mY[i];
}
}
print("Contador ", cont);
g_mT=g_mD~x1;
//
//MaxControl(-1,50);
decl vp,dfunc,ir,mhess;
vp=<1.5;1;2;4>;
ir = MaxBFGS(log_vero, &vp, &dfunc, 0, TRUE);
//
// Se convergir
if(ir==MAX_CONV){
print("\n ============================ ", j, " Replica ============================");
theta[j][0]=vp[0];
theta[j][1]=vp[1];
theta[j][2]=vp[2];
theta[j][3]=vp[3];
//
//print("\n Dados simulados ", g_mY);
//
print("\nCONVERGENCE: ",MaxConvergenceMsg(ir) );
print("\nMaximized log-likelihood: ", "%7.3f", dfunc);
print("\nML estimate: ", "%6.3f", vp);
//-------------------- Calculando os residuos ----------------------
for(i=0;i<n;i++){
decl pphi=vp[0];
decl psigma=vp[1];
decl pb0=vp[2];
decl pb1=vp[3];
decl cens=g_mT[i][1];
//
xbeta=pb0+(pb1*g_mT[i][2]);
z=(g_mT[i][0]-xbeta)/psigma;
x=log(1+exp(z));
Sy=1-probgamma(x,pphi,1);
if(cens==1){
rmi[i][j]=1+log(Sy);
if(rmi[i][j]==1){
rmi[i][j]=0.9999999999;
}
rdi[i][j]=(rmi[i][j]/fabs(rmi[i][j]))*((-2*(rmi[i][j]+log(1-rmi[i][j])))^(1/2));
}
else{
rmi[i][j]=log(Sy);
if(rmi[i][j]==0){
rmi[i][j]=-1e-5;
}
```

```
rdi[i][j]=(rmi[i][j]/fabs(rmi[i][j]))*((-(2*rmi[i][j]))^(1/2));
}
//
}
rm_ord[][j]=sortc(rmi[][j]);
rd_ord[][j]=sortc(rdi[][j]);
j++;
}
}while(j<r);
print("\n ------------------------- Resultado final ----------------------------");
//------------------- Residuos ----------------------------
decl rmResult=zeros(n,3);
decl rdResult=zeros(n,3);
for(i=0;i<n;i++){
rmResult[i][0]=min(rm_ord[i][]);
rmResult[i][1]=meanr(rm_ord[i][]);
rmResult[i][2]=max(rm_ord[i][]);
//
rdResult[i][0]=min(rd_ord[i][]);
rdResult[i][1]=meanr(rd_ord[i][]);
rdResult[i][2]=max(rd_ord[i][]);
}
//print("\n Residuo Rd ", rdResult);
//------------------- Media e variancia das r simulacoes ----------------
decl media=meanc(theta);
decl variancia=varc(theta);
decl medidas=media|variancia;
print("\nML estimate : ", media);
//------------------- Erro quadratico medio ---------------------------
decl EQM=zeros(1,4,0);
for(j=0;j<4;j++){
decl vicio=media[j]-theta0[j];
EQM[0][j]=variancia[j]+(vicio^2);
}
print("\n EQM: ", meanc(EQM));
```

References

Aarset, M. V. (1987). How to identify a bathtub hazard rate. *IEEE Transactions on Reliability, 36*, 106-108. http://dx.doi.org/10.1109/TR.1987.5222310

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

Bonat, W. H., Krainski, E. T., Ribeiro Jr, P. J., & Zeviani, W. M. (2012). *Mtodos computacional em inferncia estat stica*. Região Brasileira da Sociedade Internacional de Biometria: Piracicaba.

Carrasco, J. M. F., Ortega, E. M. M., & Paula, G. A. (2008). Log-Modified Weibull Regression Models with Censored Data: Sensitivity and Residual Analysis. *Computational Statistics and Data Analysis, 52*, 4021-4029. http://dx.doi.org/10.1016/j.csda.2008.01.027

Collett, D. (2003). *Modeling survival data in medical research*. Chapman & Hall: London.

Colosimo, E. A., & Giolo, S. R. (2006). *Anlise de sobrevivncia aplicada*. Edgard Blcher: São Paulo.

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

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. Chapman and Hall: New York.

Doornik, J. A. (2007). *An Object-Oriented Matrix Language Ox 5*. Timberlake Consultants Press: London.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics, 7*, 1-26. http://dx.doi.org/10.1214/aos/1176344552

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall: New York. http://dx.doi.org/10.1007/978-1-4899-4541-9

Ghosh, S. (2013). Normality testing for a long-memory sequence using the empirical moment generating function. *Journal of Statistical Planning and Inference, 143*, 944-954. http://dx.doi.org/10.1016/j.jspi.2012.10.016

Glänzel, W. (1987). A characterization theorem based on truncated moments and its application to some distribution families. *Mathematical Statistics and Probability Theory (Bad Tatzmannsdorf, 1986), Vol. B*. Reidel, Dordrecht, 75-84. http://dx.doi.org/10.1007/978-94-009-3965-3_8

Glänzel, W. (1990). Some consequences of a characterization theorem based on truncated moments, *Statistics: A Journal of Theoretical and Applied Statistics, 21*, 613-618.

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

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data*. John Wiley: New York. http://dx.doi.org/10.1002/9781118032985

Hashimoto, E. M., Cordeiro, G. M., & Ortega, E. M. M. (2013). The new Neyman type A beta Weibull model w ith long-term survivors.*Computational Statistics, 28*, 933-954. http://dx.doi.org/10.1007/s00180-012-0338-9

Horta, J. S. B., & Moraes, M. C. M. (2005). O sistema CAPES de avaliaão da ps-graduaão: da rea de educaão ?grande rea de cincias humanas. *Revista Brasileira de Educaão, 30*, 95-181.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. John Wiley & Sons: New Jersey.

Leiva, V., Barros, M., Paula, G. A., & Galea, M. (2007). Influence Diagnostics in Log-Birnbaum-Saunders Regression Models with Censored Data. *Computational Statistics and Data Analysis, 51*, 5694-5707. http://dx.doi.org/10.1016/j.csda.2006.09.020

Meintanis, S. G. (2010). Testing skew normality via the moment generating function. *Mathematical Methods of Statistics, 19*, 64-72. http://dx.doi.org/10.3103/S1066530710010047

Moreira, N. P., Silveira, S. F. R., Ferreira, M. A. M., & Cunha, N. R. S. (2010). Eficincia e qualidade dos programas de ps-gradua ão das instituiões federais de ensino superior usurias do programa de fomento ?ps-graduaão. *Ensaio: Avaliaão e Polticas Pblicas em Educaão, 18*, 365-388.

Mudholkar, G. S., Srivastava, D. k., & Kollia, G. D. (1996). A generalization of the Weibull distribution with application to the analysis of survival data. *Journal of the American Statistical Association, 91*, 1575-1583. http://dx.doi.org/10.1080/01621459.1996.10476725

Nadarajah, S, Cordeiro, G. M., & Ortega, E. M. M. (2015). The Zografos-Balakrishnan-G Family of Distributions: Mathematical Properties and Applications. *Communications in Statistics - Theory and Methods, 44*, 186-215.

Ortega, E. M. M., Cordeiro, G. M., & Hashimoto, E. M. (2011). A log-linear regression model for the beta-Weibull distribution. *Communication in Statistics: Simulation and Computation, 40*, 1206-1235. http://dx.doi.org/10.1080/03610918.2011.568150

Prudnikov, A. P., Brychkov, Y. A., & Marichev, O. I. (1986). *Integrals and series*. Vol. I, Taylor & Francis: London.

Ristic, M. M., & Balakrishnan, N. (2012). The gamma-exponentiated exponential distribution. *Journal of Statistical Computation and Simulation, 82*, 1191-1206. http://dx.doi.org/10.1080/00949655.2011.574633

Ross, S. M. (2006). *Simulation*. Elsevier Academic Press: Boston.

Silva, G. O., Ortega, E. M. M., Garibay, V. C., & Barreto, M. L. (2008). Log-Burr XII regression models with censored Data. *Computational Statistics and Data Analysis, 52*, 3820-3842. http://dx.doi.org/10.1016/j.csda.2008.01.003

Silva, G. O., Ortega, E. M. M., & Cordeiro, G. M. (2010). The beta modified Weibull distribution. *Lifetime Data Analysis, 16*, 409-430. http://dx.doi.org/10.1007/s10985-010-9161-1

Silva, G. O., Ortega, E. M. M., & Paula, G. A. (2011). Residuals for log-Burr XII regression models in survival analysis. *Journal of Applied Statistics, 38*, 1435-1445. http://dx.doi.org/10.1080/02664763.2010.505950

Stacy, E. W. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics, 33*, 1187-1192. http://dx.doi.org/10.1214/aoms/1177704481

Weisberg, S. (2005). *Applied linear regression*. John Wiley & Sons: New York. http://dx.doi.org/10.1002/0471704091

Xie, F. C., & Wei, B. C. (2007). Diagnostics analysis in censored generalized Poisson regression model. *Journal of Statistical Computation and Simulation, 77*, 695-708. http://dx.doi.org/10.1080/10629360600581316

Zeviani, W. M. (2012). *Estatstica computacional*. Available in:<http://www.leg.ufpr.br/ ~walmes /ensino/ce083-2012-01/ >. Access in: 30 set. 2012

Zografos, K., & Balakrishnan, N. (2009). On families of beta-and generalized gama-generated distributions and associated inference. *Statistical Methodology, 6*, 344-362. http://dx.doi.org/10.1016/j.stamet.2008.12.003

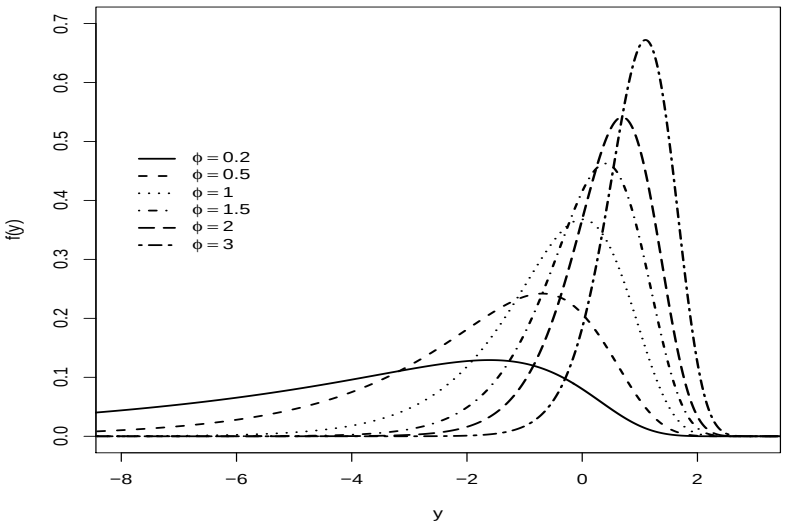

Figure 1. The log-gama-Weibull density function

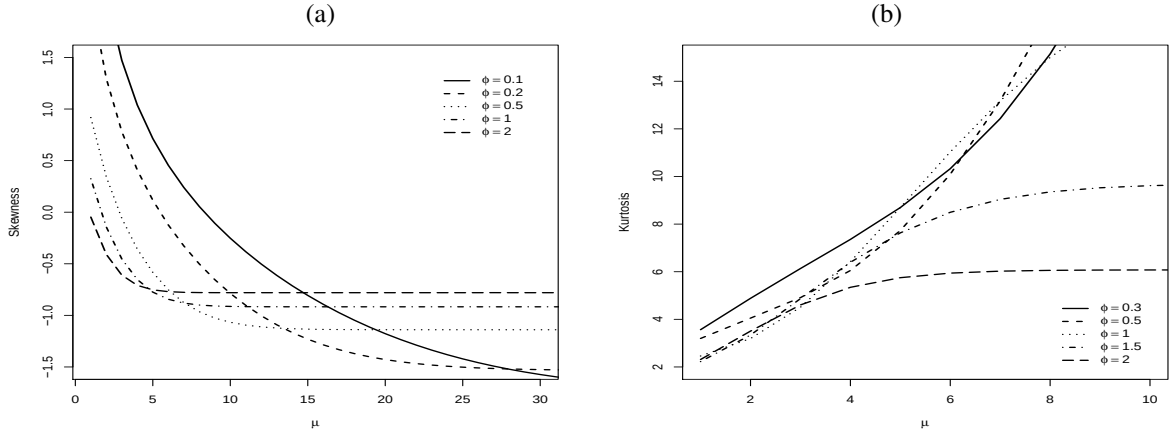

Figure 2. Skewness and kurtosis of the LGW distribution as functions of μ with $\sigma = 1.5$ and different values of ϕ.

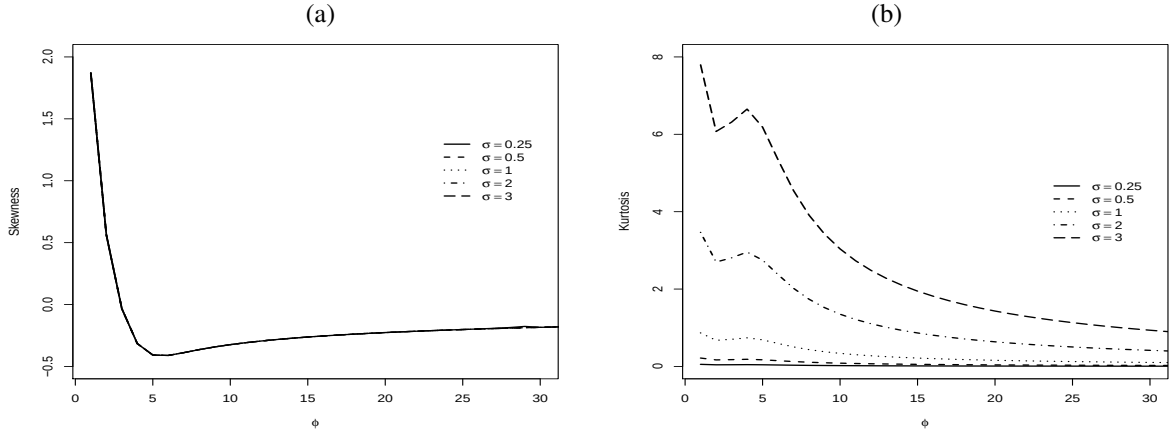

Figure 3. Skewness and kurtosis of the LGW distribution as functions of ϕ with $\mu = 0$ and different values of σ.

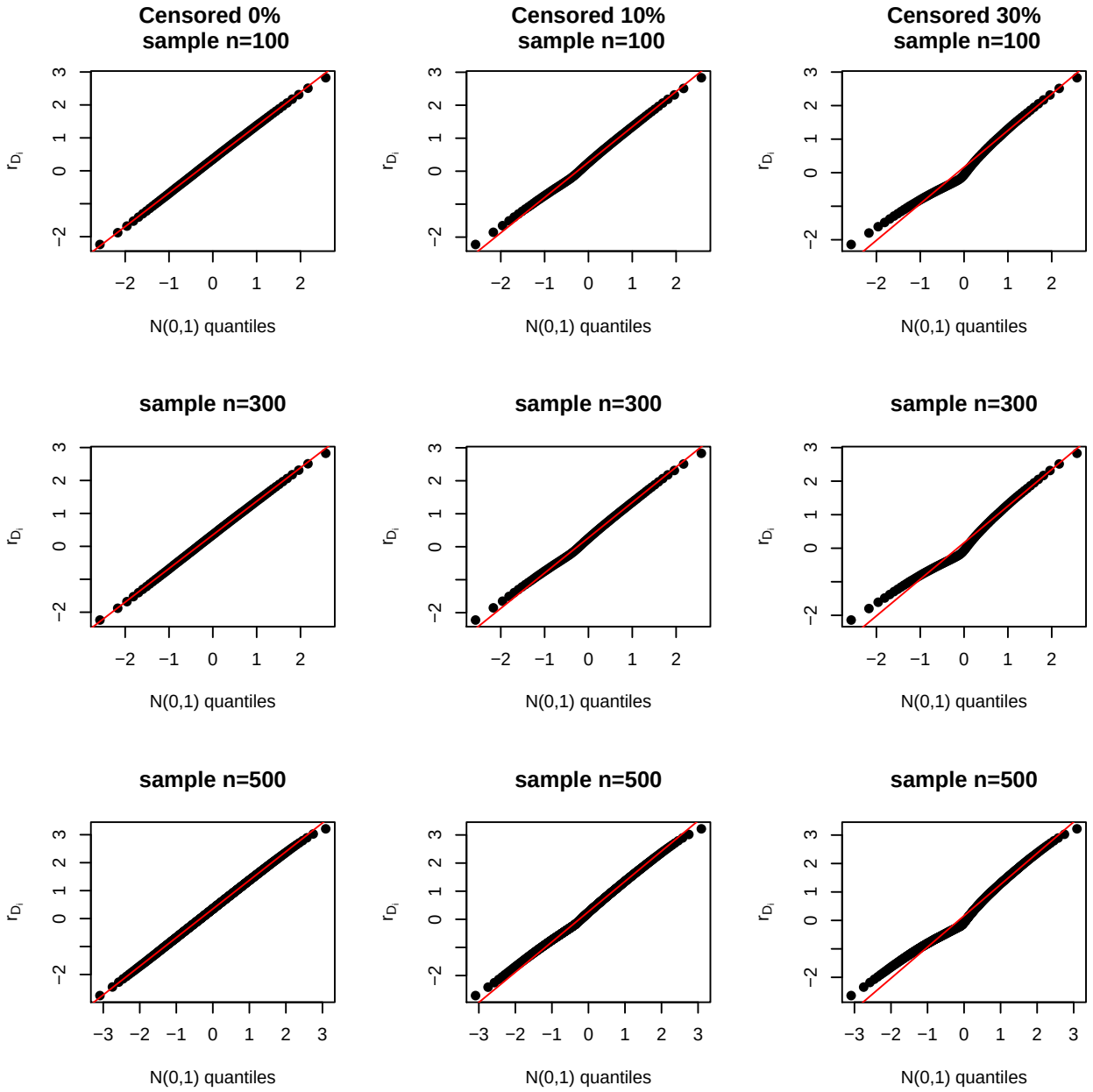

Figure 4. Normal probability plots for r_{D_i} in the LGW regression model with $\phi = 0.8$

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

	MLEs			Non-parametric bootstrap		
Parameter	Estimate	S.E.	p-value	Estimate	S.E.	95% C.I. BCa
ϕ	0.9520	1.0377	–	1.5895	0.8656	(0.2163, 1.6857)
σ	0.0630	0.0445	–	0.0801	0.0285	(0.0172, 0.0942)
β_0	4.1351	0.1590	0.0000	4.0647	0.0919	(4.0517, 4.3508)
β_1	-0.0045	0.0026	0.0884	-0.0040	0.0018	(-0.0088, -0.0023)
β_2	-0.4237	0.1112	0.0001	-0.4124	0.1061	(-0.6320, -0.2734)
β_3	0.0101	0.0030	0.0006	0.0098	0.0026	(0.0066, 0.0155)

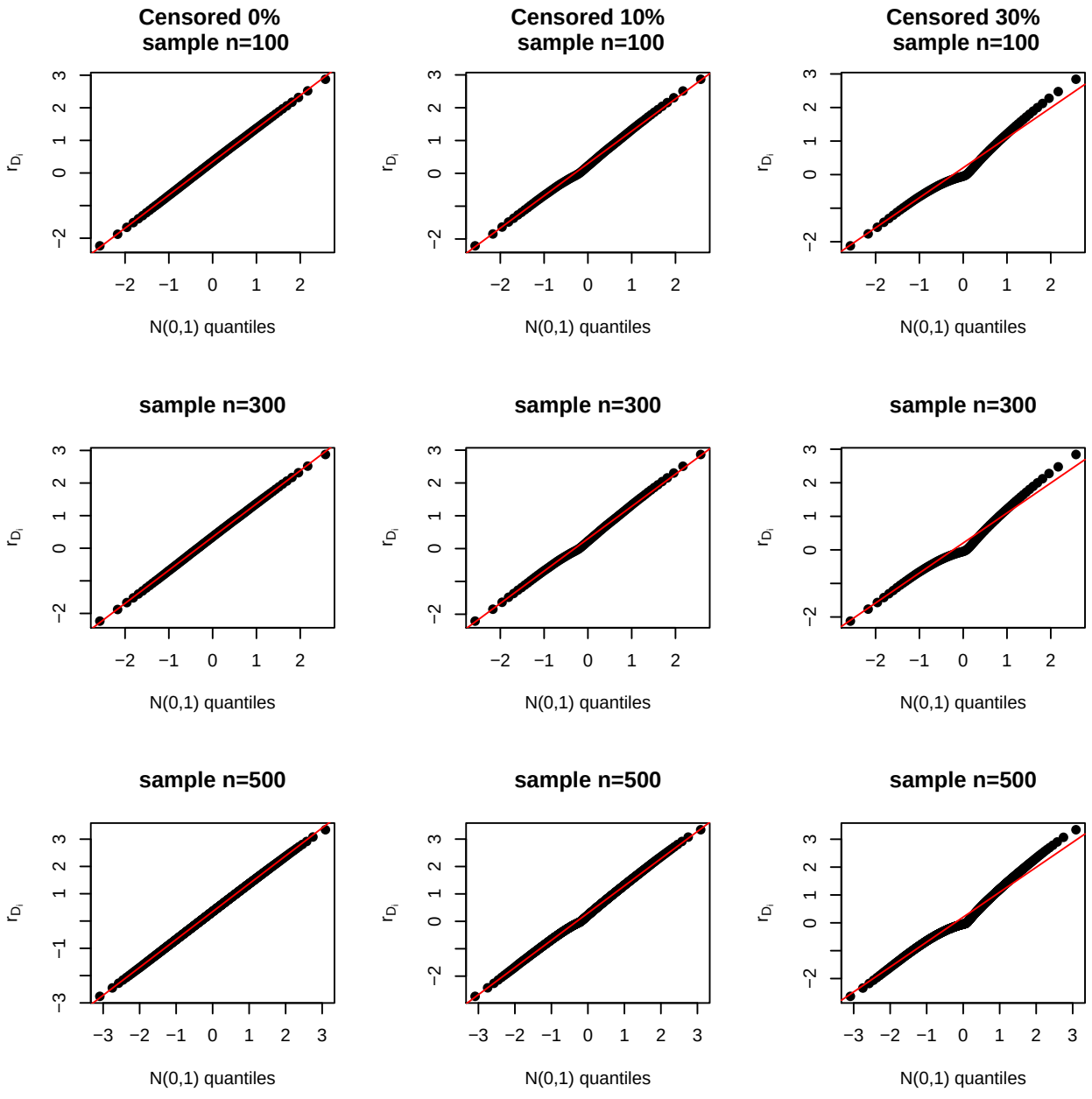

Figure 5. Normal probability plots for r_{D_i} in the LGW regression model with $\phi = 1.5$

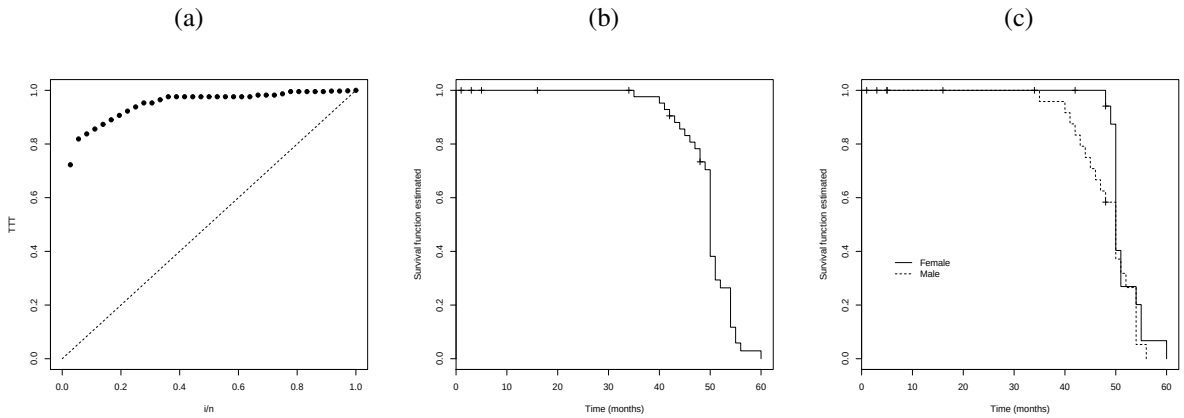

Figure 6. (a) TTT-plot. Survival curve estimate by Kaplan-Meier method for: (b) time. (c) Explanatory variable (gender)

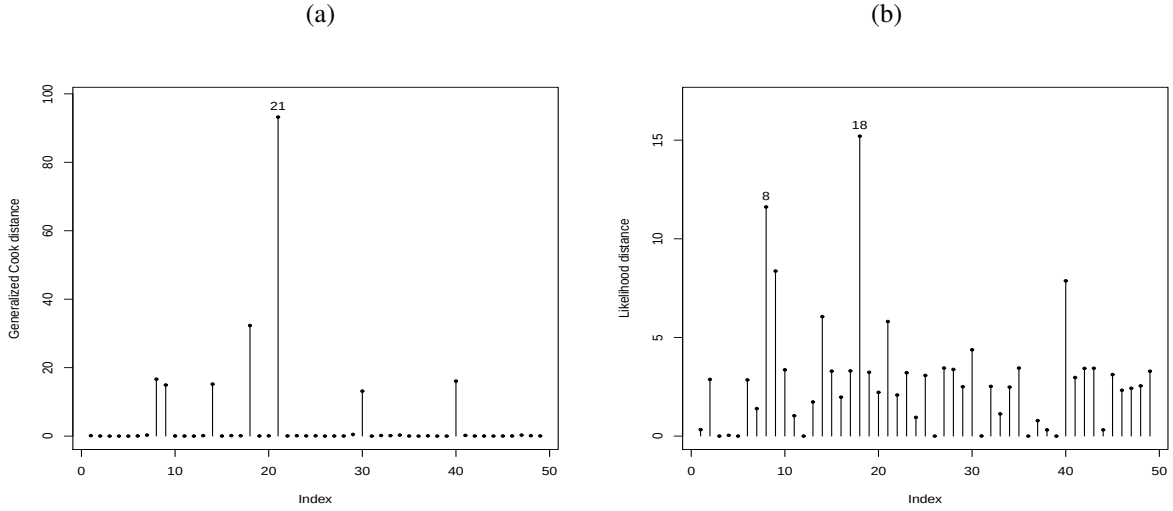

Figure 7. Index plot of global influence from the regression model fitted to the current data. (a) Generalized Cook distance. (b) Likelihood distance

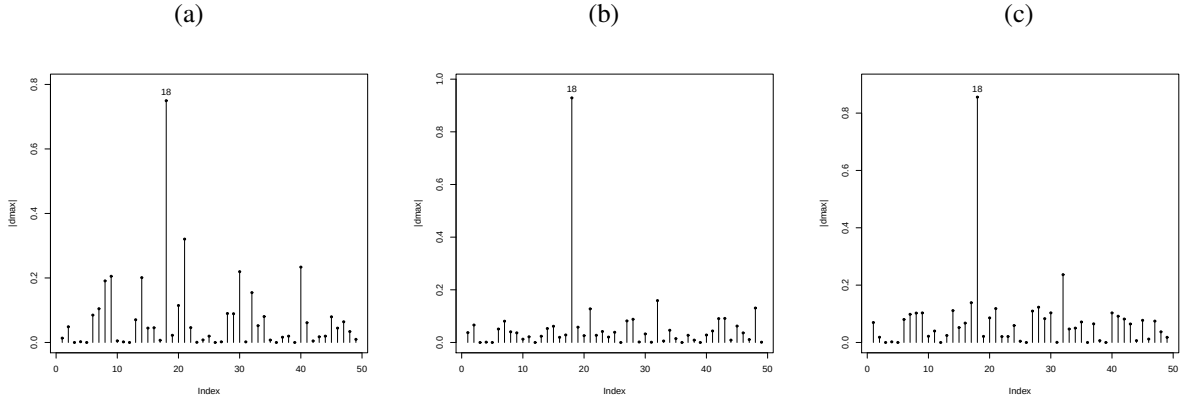

Figure 8. Index plot of $\mathbf{d}_{max}$ from the regression model fitted to the current data. (a) Case-weight perturbation. (b) Response variable perturbation. (c) Explanatory variable perturbation, x_1

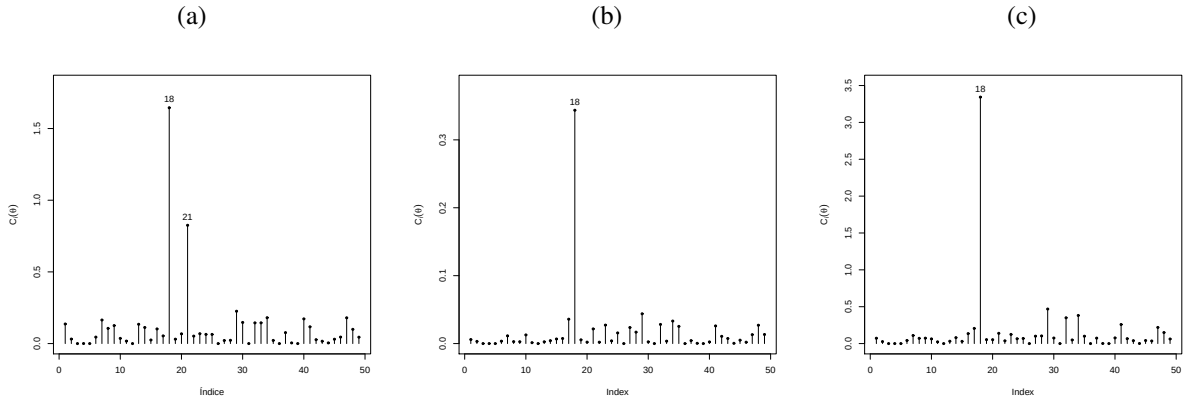

Figure 9. Index plot of local total C_i from the regression model fitted to the current data. (a) Case-weight perturbation. (b) Response variable perturbation. (c) Explanatory variable perturbation, x_1

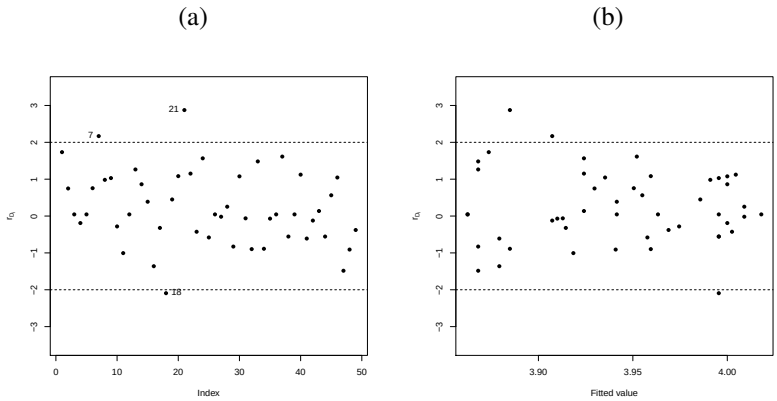

Figure 10. Residual analysis of the LGW regression model fitted to the current data. (a and b) Deviance residual

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

Set	$\hat{\phi}$	$\hat{\sigma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
A	-	-	-	-	-	-
	0.9520	0.0630	4.1351	-0.0045	-0.4237	0.0101
	(-)	(-)	(0.0000)	(0.0884)	(0.0001)	(0.0006)
A-{♯18}	[4]	[15]	[3]	[53]	[27]	[25]
	0.9257	0.0536	4.0282	-0.0021	-0.3096	0.0076
	(-)	(-)	(0.0000)	(0.2810)	(0.0007)	(0.0019)
A-{♯21}	[-351]	[-120]	[6]	[22]	[6]	[6]
	4.3450	0.1388	3.8747	-0.0035	-0.3981	0.0095
	(-)	(-)	(0.0000)	(0.1625)	(0.0004)	(0.0016)
A-{♯18, ♯21}	[49]	[51]	[2]	[51]	[31]	[29]
	0.4919	0.0310	4.0603	-0.0022	-0.2913	0.0072
	(-)	(-)	(0.0000)	(0.1892)	(0.0003)	(0.0009)

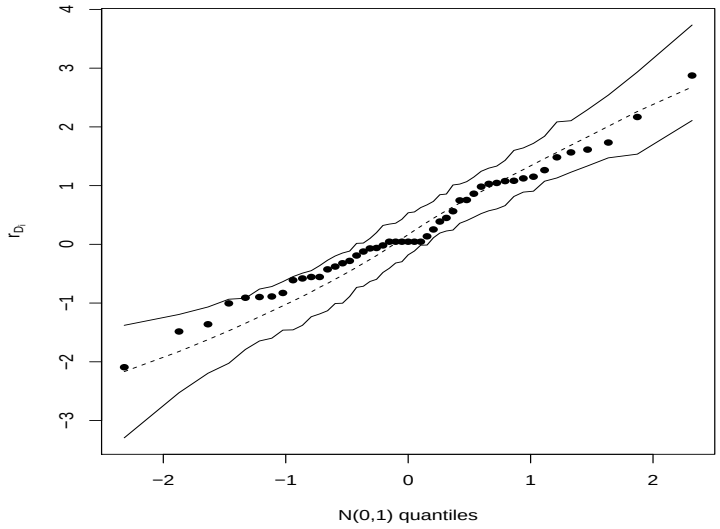

Figure 11. Normal probability plot for the deviance residual with envelopes