Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/150610

An Application of the Weibull-Poisson Long-Term Survival Regression Model

Valdemiro Piedade Vigas[1,2,*], *Giovana Oliveira Silva* [3] *and Josmar Mazucheli* [4]

[1]Departamento de Ciências Exatas, Esalq, Universidade de São Paulo, Av. Pádua Dias 11, Piracicaba, Brazil
[2]Instituto de Matemática, Universidade Federal de Mato Grosso do Sul, Av. Costa e Silva s/n, Campo Grande, Brazil
[3]Instituto de Matemática e Estatística, Universidade Federal da Bahia, Av. Adhemar de Barros s/n, Salvador, Brazil
[4]Departamento de Estatística, Universidade Estadual de Maringá, Av. Colombo 5790, Maringá, Brazil

Received: 2 May 2021, Revised: 2 Jul. 2021, Accepted: 23 Oct. 2021
Published online: 1 Nov. 2021

Abstract: In this paper, we illustrate the applicability of a regression model referring to patients with cutaneous melanoma. The reason for the choice of the dataset is due the disease be responsible for most cases of skin cancer. The proposed regression model is based on the Weibull-Poisson distribution in a structure of long-term modeling, in which the covariates were included in the proportion p of cured via the logistic link function. The motivation to study this distribution is since, in addition to generalizing the Weibull model, which is widely used in several areas of knowledge among them the survival analysis, it presents great flexibility in the forms of the hazard function. Through the proposed model we observed that the estimated proportion of cured is higher for patients with few nodules. In addition, it was presented the estimated hazard function of the patients and was verified that has a unimodal form.

Keywords: Weibull-Poisson Distribution, Long-Term Survivals Regression Model, Residual Analysis, Sensitivity Analysis.

1 Introduction

An important characteristic in the survival data arises when a part of the population is not susceptible to the event of interest, considered as cured or immune. Models in this structure have been extensively studied in the literature and are generally called long-term survival models. In this model, it is assumed that a certain proportion of individuals, say p, are cured (or immunes), in the sense that they did not present the event of interest during the period of the study. For instance, in clinical studies, a population can respond favorably to the treatment, being considered as cured. Long-term models have been used for modeling time-to-event data for various types of cancers, including breast cancer, non-Hodgkins lymphoma, leukemia, prostate cancer, and melanoma.

Berkson and Gage [1] presented a mixing model with the objective of estimate the proportion of cured in a population under treatment of stomach cancer. This model was based on a parametric distribution mixture, implying that a part of this model has an improper survival function, which represents the total of the population (cured and no cured), and a proper survival function to a part of the population, which are cured. The exponential distribution was used for the proper survival function to the proportion of cured individuals. Farewell [2] used the model proposed for Berkson and Gage, assuming the Weibull and log-normal distributions for the data of animals exposed to toxins. Yakovlev et al. [3] presented an alternative model of long-term to the Berkson and Gage model, in a structure of competing risks. Bayesian inference methods for survival data with long-term survivals were introduced by some authors such as Chen et al. [4], Hoggart and Griffin [5] and Chen and Ibrahim [6]. Besides, the authors [7,8,9,10,11,12,13] have considered this modeling approach.

Also, it is common in practical situations, that there are covariates that can influence a part of individuals cured under study. Thus, these covariates can be accommodated in the analysis in the data analysis. A model with the presence of covariates is an efficient way of to observe its effect on the proportion of cured individuals because the probability of cure of each can vary from individual to individual depending on their characteristics.

* Corresponding author e-mail: valdemiro.vigas@ufms.br

Some regression models have been proposed with this objective, among them: Ortega et al. [14] that proposed the generalized log-gamma mixture regression model; Kannan et al. [15] introduced the generalized exponential distribution cure rate model; Castro, et al. [16] defined the long-term survival models under the gamlss framework. Martinez et al. [17] used the mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data.

Recently, some authors considered the Berkson and Gage mixture model, among them we can cite: Ramires et al. [18] defined the flexible bimodal survival model with cure fraction with different regression structures; Ramires et al. [19] presented the estimation of non-linear effects in the presence of cure fraction using a semi-parametric regression model and Vigas et al. [20] proposed the Weibull regression mixture model for predicting the diarrhea data. More models in [21,22,23,24,25,26,27,28].

Thus, the objective of this paper is to propose the Weibull-Poisson long-term (LWP) regression model. This new regression model is based on Weibull-Poisson long-term distribution proposed by Vigas et al. [13] obtained by a compounding of the Weibull-Poisson (WP) distribution proposed by Bereta et al. [29] in a structure of modeling mixture proposed by Berkson and Gage [1] in which the covariates were included in the proportion of p cured via the logistic link function. The WP distribution arises by compounding of a Weibull distribution with a Poisson distribution in the context where the lifetime associated with a particular risk is not observable, instead only the minimum lifetime value among all risks is verified. However, this model can be used in any other situation as long as it fits the data satisfactorily.

In recent years, several distributions in this scenario have been proposed in the literature. For example: the exponential geometric distribution (EG) (Adamidis and Loukas [30]); the exponential Poisson distribution (EP) (Kus [31]); exponential power series distribution (EPS) (Chahkandi and Ganjali [32]); Weibull-geometric distribution (WG) (Barreto-Souza et al. [33]); complementary exponential geometric distribution (CEG) (Louzada et al. [34]); among others.

The new regression model, due to its flexibility in accommodating various forms of the risk function, (i.e., increasing, decreasing and unimodal) depending on the values of its parameters, seems to be an important model that can be applied in a variety of problems in survival data modeling. Besides, the LWP regression model is also suitable for testing goodness-of-fit of some particular sub-models, such as the exponential-Poisson and Weibull (Weibull, [35]) distributions in a structure of modeling mixture. Hence, it represents a good alternative for lifetime data analysis, and this generalization is expected to attract more comprehensive applications in survival analysis.

The inferential part of this model is carried out using the maximum likelihood approach for parameters estimation and the asymptotic distribution of the maximum likelihood estimators. Considering that the Weibull-Poisson long-term regression model is embedded in the exponential-Poisson and Weibull long-term regression models, the likelihood ratio test can be used to discriminate such models. A simulation study via Monte Carlo was conducted to evaluate the performance of the LWP regression model via bias and square root of the mean-squared error of the maximum likelihood estimates ($MLE's$).

After modeling, it is essential to check the assumptions of the model. Moreover, to conduct a robustness study to detect influential or extreme observation that can cause distortions to the results of the analysis. Numerous approaches have been proposed in the literature to detect influential or outlying observations, among them, is the global influence proposed by Cook [36]. Besides a global influence approach, was used the generalized Cook distance (Xie and Wei [37]) and the distance of likelihood (Cook and Weisberg [38]) to verify the existence of possible influential observations in the regression proposed model.

Another important step after the formulation of the model is the residuals analysis. Starting from this analysis, we can identify outliers and observe if there are differences in the assumptions made in the proposed model. In survival analysis, are proposed in the literature, some residuals for the long-term regression models. See for example Castro et al. [16]. In this paper, we considered the randomized quantile residuals ($\hat{r}$) proposed by Dunn and Smyth [39] beyond of the *Worm-Plot* an QQ-plot graphics. The graphic *Worm-Plot* introduced by Buuren and Fredriks [40] are based on the randomized quantile residuals.

This paper is organized as follows. In Sections 2 and 3, we introduce the LWP distribution and the LWP regression model, respectively. The inferential procedure based on the maximum likelihood approach; the criteria *AIC* and the likelihood ratio test to select the best model respectively is presented in Section 4. In Section 5, we present the results of a simulation study conducted to assess the performance of the maximum likelihood estimators of the new regression model. In Sections 6 and 7, the diagnostic measures and of the residual analysis are presented, respectively. In Section 8, the data set is analyzed, and the final considerations appear in Section 9.

2 The Weibull-Poisson long-term distribution (LWP)

The Weibull-Poisson long-term distribution proposed by Vigas et al. [13] obtained by a compounding of the Weibull-Poisson (WP) distribution proposed by Bereta et al. [29] in a structure of mixture modeling proposed by Berkson and Gage [1]. In this scenario, T is a

non-negative random variable that represents the lifetime in a population in which it is considered to exist cured individuals (not susceptible) with probability p and not cured individuals (susceptible) with probability (1-p). In this way, the survival function population is defined as:

$$S(t) = S_{pop}(t) = p + (1-p)S_0(t), \tag{1}$$

where $S_0(t)$ is the survival function for individuals who are susceptible. In this work, we used the survival function of the Weibull-Poisson (WP) distribution (Bereta et al. [29]) given by:

$$S_0(t) = \frac{\exp\left\{\alpha \exp\left[-(\lambda t)^{\gamma}\right]\right\} - 1}{\exp(\alpha) - 1}. \tag{2}$$

The WP model is suitable for lifetime data modeling, which has a decreasing, increasing and unimodal hazard function and it is presents as particular case some distributions used in the area, including the Weibull distribution. Replacing (2) of (1) the $S(t)$, can be written as:

$$S(t) = \frac{p\exp(\alpha) + (\exp\left\{\alpha \exp\left[-(\lambda t)^{\gamma}\right]\right\}) - 1}{\exp(\alpha) - 1} - \frac{p(\exp\left\{\alpha \exp\left[-(\lambda t)^{\gamma}\right]\right\})}{\exp(\alpha) - 1},$$

where $t > 0$; $\gamma > 0$ and $\lambda > 0$ are shape parameters; $\alpha > 0$ and $0 < p < 1$ are scale parameters. The density, hazard and quantile functions of LWP distribution are given by, respectively,

$$f(t) = \frac{(1-p)\alpha \exp\left\{\alpha \exp\left[-(\lambda t)^{\gamma}\right] - (\lambda t)^{\gamma}\right\} \lambda^{\gamma} t^{\gamma-1} \gamma}{\exp(\alpha) - 1},$$

$$h(t) = \frac{p\exp(\alpha) + (\exp\left\{\alpha \exp\left[-(\lambda t)^{\gamma}\right]\right\})}{(1-p)\alpha \exp\left\{\alpha \exp\left[-(\lambda t)^{\gamma}\right] - (\lambda t)^{\gamma}\right\} \lambda^{\gamma} t^{\gamma-1} \gamma} - \frac{p(\exp\left\{\alpha \exp\left[-(\lambda t)^{\gamma}\right]\right\}) - 1}{(1-p)\alpha \exp\left\{\alpha \exp\left[-(\lambda t)^{\gamma}\right] - (\lambda t)^{\gamma}\right\} \lambda^{\gamma} t^{\gamma-1} \gamma},$$

and

$$t = \frac{1}{\beta}\left[\log(\alpha) - \log\left(\log(\exp(\alpha)(1-p-u) + u)\right)\right]^{1/\gamma} - \frac{1}{\beta}\left[\log(1-p)\right]^{1/\gamma}. \tag{3}$$

The quantile function (3) has tractable properties specially for simulations.

Figures 1 and 2 illustrates some of the possible shapes of the hazard function according to the selected parameter values of the LWP distribution. We note from this figure that the population hazard function is quite flexible and can accommodate various forms, such as decreasing, increasing and unimodal. Applications of the LWP distribution in survival studies were investigated by Vigas et al. [13].

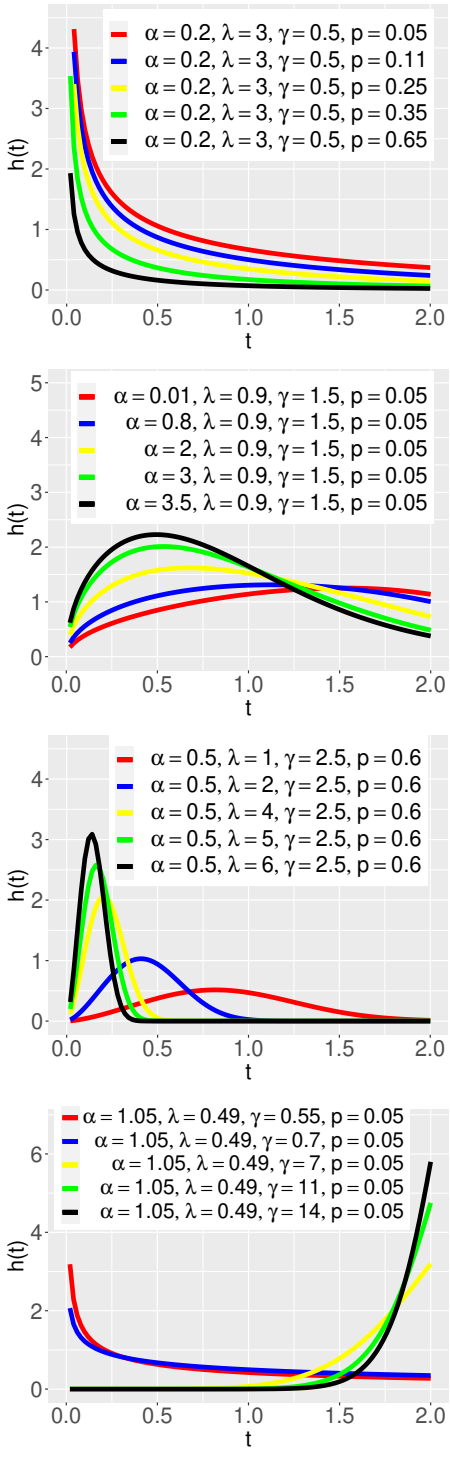


Fig. 1: Plots of the failure rate function for LWP distribution.

3 The LWP Regression Model

In many practical applications, some characteristics influence the proportion of cured; these characteristics are

called of the covariates. Thus, the model with these covariates is an efficient way to observe their effect on the proportion of cured. The covariates vector is denoted by $\mathbf{x} = (x_1, x_2, ..., x_p)^\top$, in that this vector is related to the proportion of cured of LWP model by the logistic link function. Then, the probability of cure for the i_{th} cured individual is given by

$$p(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \beta)}{1+\exp(\mathbf{x}_i^\top \beta)}, i = 1, \ldots, n,$$

where the logistic link function keeps each $p(\mathbf{x}_i)$ strictly between 0 and 1, $x_{0i} = 1$ and $\beta = (\beta_0, \beta_1, ..., \beta_k)^\top$ is the vector of unknown parameters to be estimated.

Now, consider that the covariates vector of the matrix $\mathbf{x}$ is included in the cured proportion of the variable T that follows a LWP distribution with the parameters vector $\theta = (\alpha, \lambda, \gamma, \beta^\top)^\top$. The survival and density functions are specified by, respectively:

$$S_{(t|\mathbf{x})} = \frac{p(\mathbf{x})\exp(\alpha) + (\exp\{\alpha\exp[-(\lambda t)^\gamma]\}) - 1}{\exp(\alpha) - 1} - \frac{p(\mathbf{x})(\exp\{\alpha\exp[-(\lambda t)^\gamma]\})}{\exp(\alpha) - 1},$$

and

$$f_{(t|\mathbf{x})} = \frac{(1-p(\mathbf{x}))\alpha\exp\{\alpha\exp[-(\lambda t)^\gamma] - (\lambda t)^\gamma\}\lambda^\gamma t^{\gamma-1}\gamma}{\exp(\alpha) - 1} \quad (4)$$

The LWP regression model (4) presents sub-models as particular cases because it generalizes some known distributions in the literature. It is observed that when $\alpha \to 0$ in equation (4), the LWP regression model is reduced for the Weibull long-term (LW) regression model. For $\gamma = 1$ in equation (4), the LWP regression model is reduced for the exponential-Poisson long-term (LEP) regression model.

4 Inference

4.1 Estimation by maximum likelihood

Several methods can be used to estimate the parameters of the probabilistic models, and the most common of theses methods is the maximum likelihood. One of the characteristics of this method is which allows the inclusion of censoring in its estimation process, which is not always possible with other estimation methods, for example, the least-squares method. Consider a random sample of size n composed by $(t_1, \mathbf{x}_1), (t_2, \mathbf{x}_2), \ldots, (t_n, \mathbf{x}_n)$, where t_i is the survival time with the probability density function (4) where $\mathbf{x}_i$ is the covariate vector associated with the i_{th} individual. Given that the variables T and censoring (C) are independent, the lifetimes are identically distributed and censoring is not informative. Thus, the log-likelihood function of the parameter vector θ, $\ell(\theta)$ can be written as

$$\ell(\theta) \propto \sum_{i=1}^{n}\delta_i \log\left[\alpha\gamma\lambda^\gamma t_i^{\gamma-1}\{1-(p(\mathbf{x}_i))\}\right] + \sum_{i=1}^{n}\delta_i\left[\alpha\exp\{-(\lambda t_i)^\gamma\} - (\lambda t_i)^\gamma\right] - \sum_{i=1}^{n}\delta_i\log\exp(\alpha) - 1 + \sum_{i=1}^{n}(1-\delta_i)\log\{-\exp(\alpha)\left(\exp\{\alpha\exp\left[-(\lambda t)^\gamma\right]\}\right)[(p(\mathbf{x}_i))]\} - \sum_{i=1}^{n}(1-\delta_i)\log\{\exp(\alpha) - 1\}, \quad (5)$$

where $p(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}$ and δ_i an failure indicator variable, respectively. Maximum likelihood estimates ($MLE's$) of the $\hat{\theta}$ for the parameter vector $\theta = (\alpha, \lambda, \gamma, \beta^T)^T$ is obtained maximizing (5), solving the system of equations given by

$$U(\theta) = \frac{\partial l(\theta)}{\partial\theta} = \mathbf{0}.$$

The vector $U(\theta)$ of the LWP regression model, where

$$U(\theta) = U(\alpha, \lambda, \gamma, \beta_j) = \left(\frac{\partial\ell(\theta)}{\partial\alpha}; \frac{\partial\ell(\theta)}{\partial\lambda}; \frac{\partial\ell(\theta)}{\partial\gamma}; \frac{\partial\ell(\theta)}{\partial\beta_j}\right);$$

for $j = 1, 2, \ldots, p$, has components expressed in the form

$$\frac{\partial l(\theta)}{\partial\alpha} = \frac{\sum_{i=1}^n\delta_i}{\alpha} + \sum_{i=1}^{n}\delta_i\left(\exp\{\alpha\exp\left[-(\lambda t_i)^\gamma\right]\}\right) - \frac{\sum_{i=1}^n\delta_i\exp(\alpha)}{\sum_{i=1}^n(1-\delta_i)\exp(\alpha) - 1} + \frac{\exp\{\alpha\exp\left[-(\lambda t_i)^\gamma\right] - (\lambda t_i)^\gamma\}\lambda^\gamma t_i^{\gamma-1}\{1-p(\mathbf{x}_i)\}p(\mathbf{x}_i)\exp(\alpha)}{\exp\{\alpha\exp\left[-(\beta t_i)^\gamma\right]\}(1-p(\mathbf{x}_i))p(\mathbf{x}_i)\{\exp(\alpha) - 1\}} + \frac{\exp\{-(\lambda t_i)^\gamma\}(1-p(\mathbf{x}_i))p(\mathbf{x}_i)\exp(\alpha)}{\exp\{\alpha\exp\left[-(\beta t_i)^\gamma\right]\}(1-p(\mathbf{x}_i))p(\mathbf{x}_i)\{\exp(\alpha) - 1\}} + \frac{\sum_{i=1}^n(1-\delta_i)}{\exp(\alpha) - 1}.$$

$$\frac{\partial l(\theta)}{\partial\lambda} = \frac{\sum_{i=1}^n\delta_i\gamma}{\lambda} + \sum_{i=1}^{n}\delta_i\left(\frac{-\alpha(\lambda t_i)^\gamma\gamma\exp\{-(\lambda t_i)^\gamma\}}{\lambda} - \frac{(\lambda t_i)^\gamma\gamma}{\lambda}\right) - \frac{\sum_{i=1}^n(1-\delta_i)\alpha(\lambda t_i)^\gamma\gamma\exp\{-(\lambda t_i)^\gamma\}\exp\{\alpha\exp\left[-(\lambda t_i)^\gamma\right]\}}{\lambda[\exp\{\alpha\exp\left[-(\lambda t_i)^\gamma\right]\}(1-p(\mathbf{x}_i))p(\mathbf{x}_i)\{\exp(\alpha) - 1\}]}X \frac{(1-p(\mathbf{x}_i))p(\mathbf{x}_i)\exp(\alpha)}{\lambda[\exp\{\alpha\exp\left[-(\lambda t_i)^\gamma\right]\}(1-p(\mathbf{x}_i))p(\mathbf{x}_i)\{\exp(\alpha) - 1\}]}.$$

$$\frac{\partial l(\theta)}{\partial\gamma} = \frac{\sum_{i=1}^n\delta_i[\alpha\lambda^\gamma(t_i)^{\gamma-1}(1-p(\mathbf{x}_i))}{\alpha\gamma\lambda^\gamma t_i^{\gamma-1}(1-p(\mathbf{x}_i))} + \frac{\alpha\gamma\lambda^\gamma\log(\lambda)t_i^{\gamma-1}(1-p(\mathbf{x}_i)) + \alpha\gamma\lambda^\gamma t_i^{\gamma-1}\log(t_i)(1-p(\mathbf{x}_i))]}{\alpha\gamma\lambda^\gamma t_i^{\gamma-1}(1-p(\mathbf{x}_i))} + \sum_{i=1}^{n}\delta_i\left(-\alpha(\lambda t_i)^\gamma\log(\lambda t_i)\exp\{-(\lambda t_i)^\gamma\} - (\lambda t_i)^\gamma\log(\lambda t_i)\right) - \frac{\sum_{i=1}^n(1-\delta_i)\alpha(\lambda t_i)^\gamma\log(\lambda t_i)\exp\{-(\lambda t_i)^\gamma\}}{\exp\{\alpha\exp\left[-(\lambda t_i)^\gamma\right]\}(1-p(\mathbf{x}_i))p(\mathbf{x}_i)(\exp(\alpha) - 1)}X \frac{\exp\{\alpha\exp[-(\lambda t_i)^\gamma]\}(1-p(\mathbf{x}_i))p(\mathbf{x}_i)\exp(\alpha)}{\exp\{\alpha\exp\left[-(\lambda t_i)^\gamma\right]\}(1-p(\mathbf{x}_i))p(\mathbf{x}_i)(\exp(\alpha) - 1)}.$$

$$\frac{\partial \ell(\theta)}{\partial \beta_j} = \sum_{i=1}^{n}(\delta_i)\frac{\alpha\gamma\lambda^{\gamma}\{v_i x_i(1+v_i) - v_i\{v_i x_i\}\}}{(1+v_i)^2} + \sum_{i=1}^{n}(1-\delta_i)\frac{\{\exp(\alpha)\exp\alpha\exp(-\lambda t_i^{\gamma})\}\{v_i x_i(1+v_i) - v_i(v_i x_i)\}}{(1+v_i)^2},$$

where $p(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}$ and $v_i = \exp(\mathbf{x}_i^T\beta)$.

Since the equation system to find these estimators is not linear, we can use some numerical methods for solving the system of equations. Hence, the estimates of these parameters were obtained via numerical methods. We have used the method `BFGS` through the command `optim` of the Software *R*. When the sample size is large and under certain regularity conditions for the likelihood function, confidence intervals and hypothesis testing for the parameters can be obtained using the fact that maximum likelihood estimators, $\hat{\theta}$, have asymptotic multivariate normal distribution with mean θ and variance and covariance matrix Σ, estimated by $I^{-1}(\theta) = -E[\ddot{L}(\theta)]$, where $\ddot{L}(\theta) = \left\{\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}\right\}$, this is, $\sqrt{n}(\widehat{\theta} - \theta) \sim N_{k+3}(\mathbf{0}, \mathbf{I}^{-1}(\theta))$. Whereas the calculation of the $\mathbf{I}(\theta)$ is not possible by the presence of censored observations, alternatively we can use the information matrix observed, -$\ddot{L}(\theta)$, assessed in $\theta = \widehat{\theta}$, which is a consistent estimator for Σ. For more details, see [41], [42] and [43].

4.2 Model Selection

As the LWP regression model is reduced in sub-models, the selection criteria may be used to choose the more appropriate model. The *AIC* (Akaike's information criterion) and the likelihood ratio test was used for the selection of the model that best fits the data. The *AIC* criterion is defined by $AIC = -2\log(\hat{L}) + 2k$, where: $\hat{L}$ is the maximized value of the likelihood function of the model; k is the number of parameters of the model; and n is the sample size. The preferred model is the one with the smallest value of *AIC*. Besides this criterion, hypothesis tests, such as the likelihood ratio test (LR), can be taken into account due the LWP distribution has other distributions as particular cases.

The likelihood ratio test (LR) is used to discriminate nested models. To test nested distributions, we compute the maximum values of the restricted (H_0) and unrestricted (H_1) log-likelihoods to construct the test statistic. Under H_0 and some regularity conditions, the distribution of the statistical likelihood ratio (ω_n) converges to a χ^2 distribution with degrees of freedom equal to the difference between the numbers of parameters of the unrestricted and restricted models.

For example, hypotheses are given by $H_0 : \gamma = 1$ (LEP model) versus $H_1 : \gamma \neq 1$ (LWP model). The test statistic is given by $\omega_n = -2 \times \log\left(\frac{L(\widehat{\mathbf{I}(\theta)}_0)}{L(\widehat{\mathbf{I}(\theta)})}\right)$, where $\widehat{\theta}_0$ is the maximum likelihood estimator for θ under H_0, and the null hypothesis is rejected when $\omega_n > \chi^2_{1-\alpha}(1)$, which is the quantile of the chi-square distribution with one degree of freedom.

For the comparison of the models in the boundary of the parameter space, for example, $H_0 : \alpha \to 0$ and $H_1 : \alpha > 0$, the distribution of the statistical test ω_n is a mixture with a weight (0.5 and 0.5) of distribution χ^2 with one degree of freedom, a discrete distribution and concentrated mass in the value 0, this is, $P(\omega_n \leq w) = \frac{1}{2} + \frac{1}{2}P(\chi^2_1 \leq w)$. Large positive values of ω_n give favorable evidence to the unrestricted model. For example, for a significance level of 5%, H_0 is rejected if $\omega_n > 2.7055$. More details in [7] and [44].

5 Simulation Study

To assess the performance of *MLE*$'s$ for the parameters of the LWP regression model, a simulation study was made for different values of n. In this study, the survival time of T follows WP distribution for different values of n (200, 400, 600 and 800) in which the covariate was included in the proportion of p cured via the logistic link function. The values of the WP variable were generated from the inverse transformation method. Through this method, we can obtain, in a closed form, generate values of the LWP distribution via the quantile function given by equation (3). The censoring times C were sampled from the Uniform distribution in the $(0,\tau)$ interval where the τ controls the censoring observations. Besides, the survival time of the variable (Y) in the simulation was considered through $y_i = \min(t_i, c_i)$.

The parameter values were fixed in $\alpha = 3.5$, $\lambda = 0.5$, and $\gamma = 1.5$ with different values of τ (4, 5, 6 and 7) for each analyzed sample of the LWP, LEP and LW regression models, where $p_i = \exp(\beta_0 + \beta_1 x_i)/(1+\exp(\beta_0 + \beta_1 x_i))$. The values of the covariate x_i from Bernoulli distribution with parameter 0.5, considering $\beta_0 = 0.4$ and $\beta_1 = -0.5$ with the proportion of cured for the two levels of x (0,1), were $p^{(x=0)} = 0.5986 = p_0$ and $p^{(x=1)} = 0.4750 = p_1$ respectively. The process of this simulation is as follows:

1. Generate $M_i \sim$ Bernoulli (p_i);
2. If $M_i = 0$, then $t_i = \infty$, otherwise generate $T \sim$ WP$(\alpha, \lambda, \gamma)$;
3. Generate variable of censure c_i; $C \sim$ U$(0;\tau)$;
4. Find $y_i = \min(t_i, c_i)$;
5. If $y_i < c_i$ then $\delta_i = 1$, otherwise, $\delta_i = 0$, to $i = 1, ..., n$.

For each combination of n, 1000 samples were generated and were obtained the maximum likelihood estimates of the LWP regression model. The bias and the square root of the mean-squared error (*RMSE*) of the maximum likelihood estimates were also calculated for simulated samples in the same conditions of the previous simulations. From the simulation results, shown in Tables 1, 2, 3 and 4, it was observed that the estimates of the parameters of the LWP regression model were close to the true value of the parameters and the *RMSEs* decay toward zero when the sample size n increases as expected. These facts support that the asymptotic normal distribution provides an adequate approximation to the finite sample of the estimators distribution. The affirmations about the maximum likelihood estimates and *RMSEs* remain valid when we increase the values of the parameter τ. Table 5 shows the estimated proportion of cured for the two levels of x, and it can be observed that this proportion is close to the true value of the parameters for the two levels of x. For all sample sizes, the convergence rate was calculated for all scenarios, and it was found that the rate was 100% in all cases in this work.

Table 1: AEs, biases and RMSEs for the parameters of LWP regression model for different values of $\alpha, \lambda, \gamma, \beta_0, \beta_1, n = 200$

	$\tau = 4$			$\tau = 5$		
θ	AE	Bias	RMSE	AE	Bias	RMSE
α	4.4880	0.9880	2.8557	4.2769	0.7769	2.7203
λ	0.5102	0.0102	0.2108	0.5168	0.0168	0.2220
γ	1.5157	0.0157	0.1562	1.5039	0.0039	0.1430
β_0	0.4109	0.0109	0.2703	0.4093	0.0093	0.2518
β_1	-0.5286	-0.0286	0.3386	-0.5217	-0.0217	0.3245
	$\tau = 6$			$\tau = 7$		
θ	AE	Bias	RMSE	AE	Bias	RMSE
α	3.8390	0.3390	2.5027	3.7740	0.2740	2.5167
λ	0.5502	0.0502	0.2283	0.5565	0.0565	0.2374
γ	1.4983	-0.0016	0.1377	1.4922	-0.0077	0.1361
β_0	0.3905	0.0094	0.2286	0.3978	-0.0021	0.2244
β_1	-0.5032	-0.0032	0.2965	-0.4974	0.0025	0.3010

Table 2: AEs, biases and RMSEs for the parameters of LWP regression model for different values of $\alpha, \lambda, \gamma, \beta_0, \beta_1, n = 400$

	$\tau = 4$			$\tau = 5$		
θ	AE	Bias	RMSE	AE	Bias	RMSE
α	4.5941	1.0941	2.7779	4.3022	0.8022	2.5938
λ	0.4878	-0.0121	0.1964	0.5052	0.0052	0.1962
γ	1.4961	-0.0262	0.1085	1.4940	-0.0059	0.1033
β_0	0.3924	-0.0038	0.1854	0.3963	-0.0063	0.1733
β_1	-0.5101	-0.0101	0.2434	-0.5126	-0.0012	0.2340
	$\tau = 6$			$\tau = 7$		
θ	AE	Bias	RMSE	AE	Bias	RMSE
α	3.8793	0.3793	2.3733	3.6384	0.1984	2.2599
λ	0.5341	0.034	0.2031	0.5485	0.0485	0.2053
γ	1.4990	-0.0341	0.1006	1.4922	-0.0077	0.0956
β_0	0.3976	-0.0023	0.1654	0.4005	0.0005	0.1622
β_1	-0.4998	0.0001	0.2256	-0.5113	-0.0113	0.2215

Table 3: AEs, biases and RMSEs for the parameters of LWP regression model for different values of $\alpha, \lambda, \gamma, \beta_0, \beta_1, n = 600$

	$\tau = 4$			$\tau = 5$		
θ	AE	Bias	RMSE	AE	Bias	RMSE
α	4.5584	1.0584	2.6905	4.3114	0.8114	2.5227
λ	0.4804	-0.0195	0.1833	0.4955	-0.0044	0.1848
γ	1.4930	-0.0069	0.0936	1.4901	-0.0098	0.0892
β_0	0.3982	-0.0017	0.1564	0.3976	-0.0023	0.1428
β_1	-0.5066	-0.0066	0.2039	-0.5022	-0.0022	0.1959
	$\tau = 6$			$\tau = 7$		
θ	AE	Bias	RMSE	AE	Bias	RMSE
α	3.8667	0.3667	2.2051	3.8661	0.3681	2.2149
λ	0.5246	0.0246	0.1828	0.5248	0.0248	0.1837
γ	1.4903	-0.0096	0.0873	1.4938	-0.0061	0.0846
β_0	0.4015	-0.0015	0.1355	0.4027	0.0027	0.1361
β_1	-0.5005	-0.0055	0.1841	-0.5023	-0.0023	0.1900

Table 4: AEs, biases and RMSEs for the parameters of LWP regression model for different values of $\alpha, \lambda, \gamma, \beta_0, \beta_1, n = 800$

	$\tau = 4$			$\tau = 5$		
	$n = 800$			$n = 800$		
θ	AE	Bias	RMSE	AE	Bias	RMSE
α	4.5182	1.0182	2.6539	4.1748	0.6748	2.2979
λ	0.4847	-0.0152	0.1792	0.4958	-0.0041	0.1711
γ	1.4946	-0.0053	0.0800	1.4878	-0.0121	0.0798
β_0	0.3853	-0.0146	0.1302	0.3957	-0.0042	0.1268
β_1	-0.4966	0.0033	0.1726	-0.4973	0.0026	0.1656
	$\tau = 6$			$\tau = 7$		
	$n = 800$			$n = 800$		
θ	AE	Bias	RMSE	AE	Bias	RMSE
α	3.9620	0.4620	2.1896	3.5940	0.0940	1.9435
λ	0.5134	0.0134	0.1725	0.5394	0.0394	0.1743
γ	1.4891	-0.0108	0.0759	1.4853	-0.0146	0.0699
β_0	0.3983	-0.0016	0.1207	0.3954	0.0045	0.1116
β_1	-0.5016	-0.0016	0.1601	-0.4990	0.0009	0.1539

Table 5: Estimated proportions of LWP regression model for different values of τ and n.

Sample size	Parameters	τ value			
		4	5	6	7
$n = 200$	p_0	0.6013	0.5954	0.5987	0.5981
	p_1	0.4706	0.4718	0.4763	0.4751
$n = 400$	p_0	0.5968	0.5978	0.5981	0.5988
	p_1	0.4706	0.4709	0.4744	0.4723
$n = 600$	p_0	0.5982	0.5978	0.5900	0.5993
	p_1	0.4729	0.4709	0.4740	0.4751
$n = 800$	p_0	0.5951	0.5976	0.5982	0.5975
	p_1	0.4721	0.4746	0.4741	0.4741

6 Sensitivity Analysis

After fitting the model, it is essential to check its assumptions and to conduct a robust study to detect influential observations that can cause distortions in the analysis. The first tool to perform sensitivity analysis is the global influence starting from case deletion (Cook [36]). Case deletion is a common approach to study the effect of dropping the i_{th} case from the data set. Another approach was suggested by (Cook and Weisberg [38]), where instead of removing observations, weights are given to them. Several authors considered this context, for example: Leiva et. al [45] who investigated the local influence in log-Birnbaum-Saunders regression model with censored observations; Silva et. al, [46], that considered the problem of assessing local influence in log-Burr regression model with censoring; and Ortega et. al [14], that discussed the local influence for the generalized log-gamma mixture model with covariates.

Global Influence

The first diagnostic measure used to evaluate the global influence in the analysis of data was Cook's distance (Cook [36]). This distance is based on the elimination of cases, which is a common approach to studying the effect of removing a case of the analysis, so it is possible to determine which individuals may influence the results of that study.

Consider an observed sample given by $(T_i, \delta_i, \mathbf{x}_i)$ to each individual i=1,2,...,n, where T_i is the survival time, $T \sim$ LWP (θ), δ_i is a censoring indicator variable and $\mathbf{x}_i$ is the vector of covariates. Now, consider that "(i)" represents that the i_{th} case was deleted. Then, the logarithm of the likelihood function of θ excluding the i_{th} individual is defined as $l_{(i)}(\theta)$ and $\hat{\theta}_{(i)} = (\alpha_{(i)}, \lambda_{(i)}, \gamma_{(i)}, \beta_{(i)}{}^T)^T$ is the maximum likelihood estimator of θ from $l_{(i)}(\theta)$. The idea to evaluate the influence of the i_{th} case of the maximum likelihood estimate $\hat{\theta}$ is to use the difference between $\hat{\theta}_{(i)}$ and $\hat{\theta}$. Then, if $\hat{\theta}_{(i)}$ is far from $\hat{\theta}$, the i_{th} case can be considered as an influential observation. A measure possible of global influence is a generalization of Cook's distance given by

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

Another measure to verify the existence of possible influential points is defined as the distance of likelihood (Cook and Weisberg [38]), which is defined by

$$LD_i(\theta) = 2\left\{l(\hat{\theta}) - l(\hat{\theta}_{(i)})\right\},$$

where $l(\hat{\theta})$ is the value of the logarithm of the likelihood function of the complete sample and $l(\hat{\theta}_{(i)})$ is the value of the logarithm of the likelihood function of the complete sample without i_{th} observation.

7 Residual Analysis

An important step after the model formulation is the analysis of residual. It is used to verify the assumptions of the model proposed, beyond detect the presence of outlying observations. In this paper, was considered the randomized quantile residuals $(\hat{r})$ proposed by Dunn [39]. These residuals are also used for regression models with censored data where the variable response is of the continuous type. The randomized quantile residuals are given by

$$\hat{r}_i = \Phi^{-1}(\hat{u}_i);\ i = 1,2,\ldots,n,$$

where Φ^{-1} corresponds the inverse of the quantile function of the standard Normal and $F(.)$ is a cumulative distribution function of the proposed model. For continuous censored variables to the right, $\hat{u}_i$ is defined in the interval $\hat{u}_i = [F(y_i|\widehat{\theta}), 1]$. The $\hat{u}_i$ to LWP regression model is defined by

$$\hat{u}_i = \frac{(1 - p(\mathbf{x}_i))\left(\exp(\alpha) - \exp\left[\alpha \exp\left(-\lambda t_i\right)^{\gamma}\right]\right)}{\exp(\alpha) - 1}$$

where $p(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}$, i=1, ..., n.

The QQ-plot of the normalized randomized quantile residuals (Paiva [47]) and the graphic *Worm-Plot* also were used to verify if the proposed model fits well to the data. The graphic *Worm-Plot* introduced by Buuren and Fredriks [40] is based on the randomized quantile residuals. If the residuals are inside of a non-rejection region (between the two elliptic curves), the global model provides a good fit. This graphic can be obtained from the function *wp* in the package GAMLSS of the Software *R*.

8 Aplication

To illustrate this new model was used the real data set of cancer recurrence. The data were part of an essay on cutaneous melanoma (a type of malign cancer) for the evaluation of postoperative treatment performance with a high dose of a certain drug (interferon alfa-2b) in order to prevent a recurrence. Patients were included in the study from 1991 to 1995, and follow up was conducted until 1998.

The data were described and analyzed by Silva et al. [48]. The original sample comprises 427 patients, but the tumor thickness data (as a covariate) of 10 patients were missing and so were removed from our analysis. Then we obtain a data set of $n = 417$ patients with approximately 56% of censored observations.

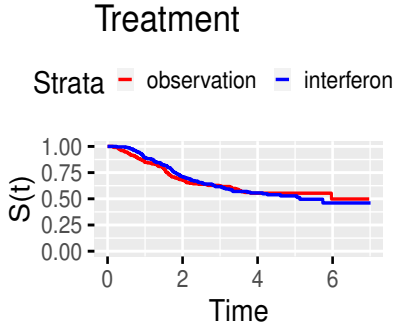
Cutaneous melanoma is the most serious type of skin cancer, as it has a high possibility of spreading to neighboring tissues and organs. In addition to men over 40 years, people with fair skin, blue eyes, blond and red hair are more susceptible to develop this disease. [1].

For each patient $i = 1,2,\ldots,417$ were registered the variable t_i: observed time (in years) until the cancer recurrence, besides of the covariates x_{i1}: treatment (0= observation, 1=interferon dose); x_{i2}: age (in years); x_{i3}: nodule (nodule category: 1 to 4); x_{i4}: gender (0=male, 1=female); x_{i5}: p.s. (performance status patients functional capacity scale as regards his daily activities: 0=fully active, 1=other); x_{i6}: tumor (tumor thickness in mm).

The aim this study is to compare the LWP regression model with the LW and LEP regression models to identify a more appropriate model for the survival time, relating to the proportion of cured patients (p) with the covariates observed. Initially, to obtain more information on the survival time, it was made an analysis of these times without considering the observations censored. This information is shown in Table 10. It can be observed following this table, that the median time of recurrence of the patients was approximate of 3.22 years, which indicates that approximately 50% of patients had the survival time larger than 3.22 years and its mean survival time was 3.17 years. It can also be observed that 25% of patients had a lifetime less than 1.67 years or greater than 4.49 years. Furthermore, the lifetime of the patients was between 0.14 and 7.01 years, which implies a greater variability of the lifetime. What certifies this greater variability is the coefficient of variation (C.V.) with the value of 53.21%. Figure 2 shows the survival estimates by Kaplan-Meier for groups of patients on the covariates: Treatment, Age, Nodule category, Sex, P.S, and Tumor. From this figure, it was noted that there is evidence of a difference between survival functions only the category of the covariate nodule (X_3).

Table 6: Summary of survival time of the patients for the cutaneous melanoma data.

Minimum	1^{o} Quart.	Median	Mean	3^{o} Quart.	Maximum	S.D.	C.V.
0.1478	1.6701	3.2224	3.1792	4.4928	7.0116	1.69170	0.5321



[1] https://https://www.rededorsaoluiz.com.br/hospital/vivalle/noticias/artigo/cancer-de-pele-melanoma-cutaneo

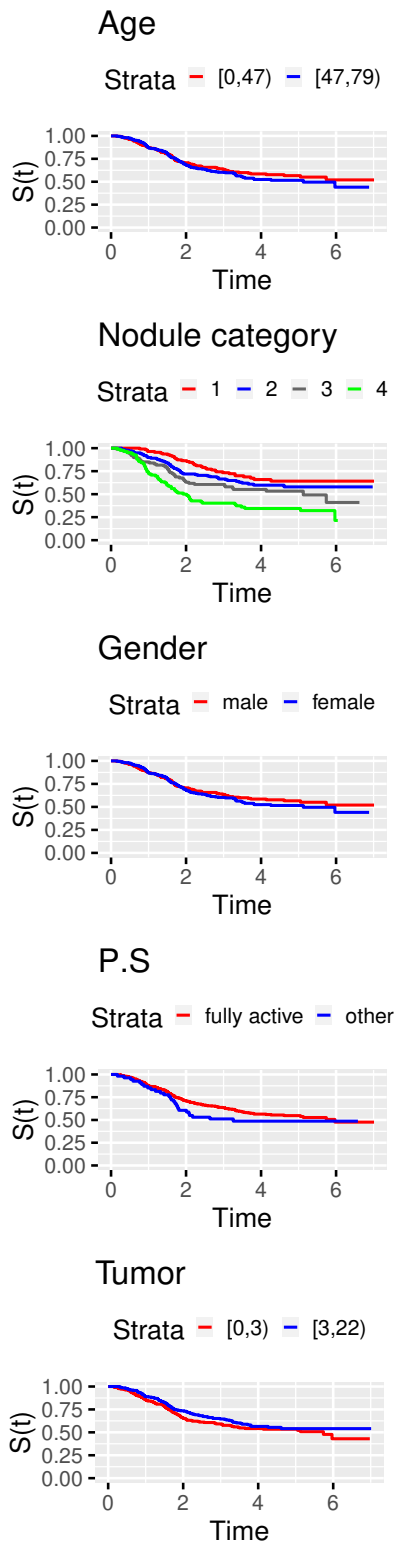


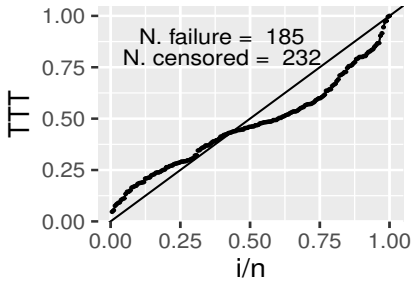
Fig. 2: Curve of the survival function estimated by the Kaplan-Meier to covariates Treatment, Age, Nodule category, Gender, P.S and Tumor.

As there is evidence of a difference between survival functions of the covariate X_3, it was done a table to verify the nodules frequency. It can be observed under the table 7 that the category "2" had the highest nodule frequency. Then, was created three dummy variable (X_{31}, X_{33} and X_{34}) in what the reference value is the nodule with the value "2".

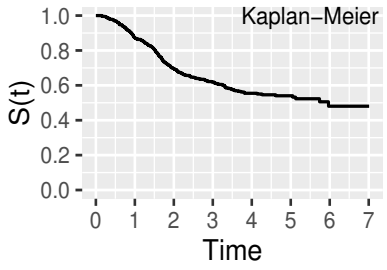
Table 7: Frequency of the variable Nodule category for the cutaneous melanoma data.

Nodule category	1	2	3	4
Frequency	111	137	87	82

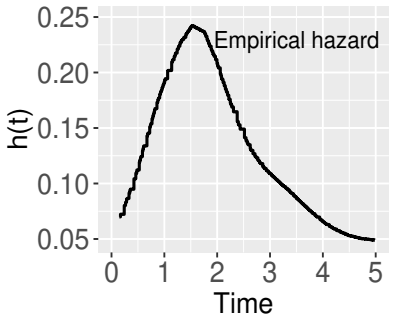
Also, it was verified the hazard function of data observed using a graphical method based on the total time test (TTT), also called TTT-Plot. This method is useful when there is information about the hazard function of the times observed. The empirical version of the TTT-Plot according to Aarset [49] is given by $G(r/n) = [(\sum_{i=1}^{r} Y_{i:n}) - (n-r)Y_{r:n}]/(\sum_{i=1}^{r} Y_{i:n})$, where $r = 1,\ldots,n$ and $Y_{i:n}$ represents the order statistics of the sample. Aarset [49] showed that: the hazard function is constant if the TTT-Plot is presented graphically as a straight diagonal; the hazard function is decreasing (or increasing) if the TTT-Plot is convex (or concave); the hazard function is U-shaped if the TTT-Plot is convex and then concave. Otherwise, the hazard function is unimodal. The TTT-Plot for the cutaneous melanoma data in Figure 3 (a) indicates a unimodal shaped failure rate function. Figure 3 (b) shows the graphic of the survival function estimated Kaplan-Meier. From this figure, it is possible to observe a fraction of cured individuals, because its estimated survival function tends to a constant value well above zero. Then, we can utilize long-term survival models and can use the LWP distribution for the modeling of data.



(a) TTT-Plot



(b) Kaplan-Meier



(c) Hazard function

Fig. 3: TTT-Plot; Kaplan Meier and empirical Hazard for the cutaneous melanoma data.

The next step after the conclusions made about the LWP model was to check which is the best model when the covariates were included. Now, we go to include the covariates vector in the proportion p of cured of LWP model by logistic function, i.e,

$$p(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}; i = 1,2,...,417,$$

where $\mathbf{x}_i^T\beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{31} x_{i31} + \beta_{33} x_{i33} + \beta_{34} x_{i34} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}$. This procedure was given in the section 3.

We compared the LWP regression model with its particular cases LEP and LW, considering the *AIC*. The values of *AIC* is presented in Table 8 with the maximum likelihood estimates ($MLE's$) and their standard errors (S.E.) for the parameters of the three models. We have used the command *optim* through of the method *BFGS* of the software *R* to calculate the $MLE's$ of the parameters of the models. It can be observed in Table 8 that the LWP regression model had the lowest value of the *AIC* concerning the LEP and LW models, indicating that this model is more appropriate to the data. For LWP and LW regression models, only the variable (X_{34}) was significant for the cure fraction. While in the LEP regression model, only the variable (X_{31}) was significant for the proportion p of cured at 5% of significance. The estimates of the parameters and their standard errors of the models were similar in most cases to LWP and LW regression models.

As the LWP regression model is reduced in the LEP and LW regression models, the likelihood ratio test was used to select a model which best fit the data. This procedure was given in section 4.2. To the LEP and LWP regression models, hypotheses were: $H_0 : \gamma = 1$, i.e, the LEP regression model is adequate versus $H_1 : \gamma \neq 1$, i.e, the LWP regression model is adequate. In this data, it was observed that the test statistic was 35.3154 (p-value < 0.001), and this result leads us to reject the null hypothesis. In relation the LW and LWP regression models, hypotheses were: $H_0 : \alpha \to 0$, i.e, the LW regression model is adequate versus $H_1 : \alpha > 0$, i.e, the LWP regression model is adequate. In this data, it was observed that the test statistic was 5.4244 bigger than $1/2 + 1/2\ P(\chi_1^2 \leq c) = 2.7055$, at the significance level of the 5%, which leads us to reject the null hypothesis. In both cases, there is evidence in favor of the LWP regression model.

Table 8: Estimated values of the parameters regression models LEP, LW and LWP for the cutaneous melanoma data..

	LEP			LW		
θ	Estimates	S.E.	p-value	Estimates	S.E.	p-value
α	7.2948	8.4322	-	-	-	-
λ	0.0340	0.0395	-	0.4495	0.0280	-
γ	-	-	-	1.6102	0.1067	-
β_0	1.5468	0.9083	0.0886	1.1150	0.4952	0.0243
β_1	-0.1721	0.4188	0.6843	-0.1556	0.2249	0.4889
β_2	-0.0297	0.0169	0.0794	-0.0140	0.0086	0.1058
β_{31}	1.6598	0.7966	0.0371	0.6332	0.3328	0.0570
β_{33}	-0.9817	0.7860	0.2116	-0.3013	0.3033	0.3204
β_{34}	-7.0650	0.6684	0.9355	-1.1032	0.3300	< 0.001
β_4	0.2986	0.4187	0.4761	0.2080	0.2324	0.3707
β_5	-0.1344	0.6486	0.8348	-0.1497	0.3368	0.6565
β_6	-0.2868	0.1659	0.0839	-0.0661	0.0424	0.1193
$-\ell(\cdot)$		528.3521			513.4066	
AIC		1078.704			1048.813	
ω_n		35.3154			5.4244	

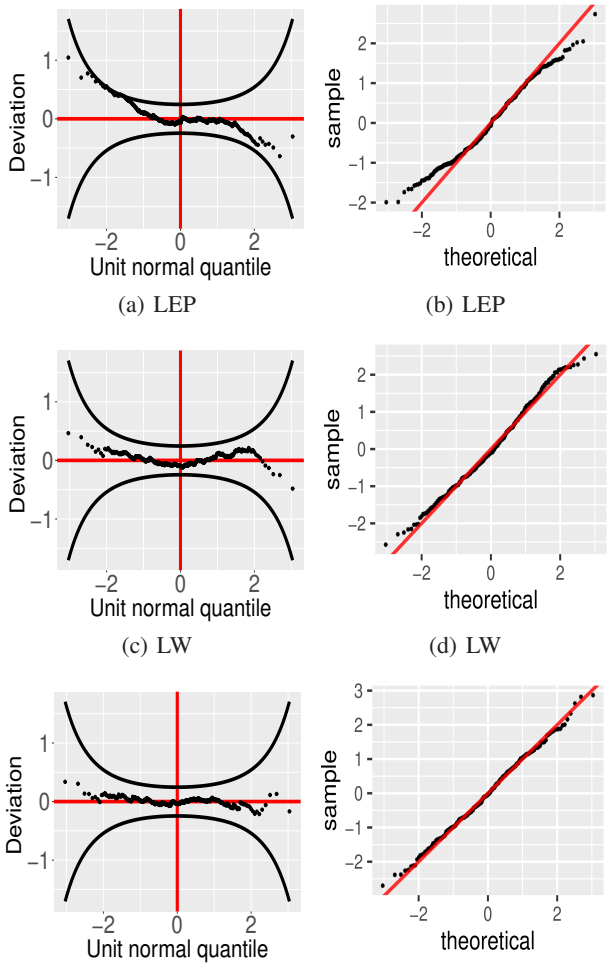
	LWP		
θ	Estimates	S.E.	p-value
α	2.8411	1.5152	-
λ	0.2698	0.0732	-
γ	1.8184	0.1341	-
β_0	1.1196	0.5024	0.0258
β_1	-0.1540	0.2284	0.5002
β_2	-0.0143	0.0088	0.1027
β_{31}	0.6471	0.3384	0.0550
β_{33}	-0.3085	0.3074	0.3155
β_{34}	-1.1329	0.3405	< 0.001
β_4	0.2050	0.2359	0.3847
β_5	-0.1480	0.3430	0.6660
β_6	-0.0661	0.0442	0.1344
$-\ell(\cdot)$		510.6944	
AIC		1045.389	
ω_n		-	

The next step after the conclusions made anteriorly about the LWP model is the residual analysis, which is useful to verify the goodness of fit of the model. Figure 4 shows the *Worm-Plot* graphic of the quantile residuals and qqnorm graphic for LEP, LW, and LWP models, respectively. The Figure 4 indicates that the LWP model is more acceptable than LEP and LW models, because the graphic *Worm-Plot* of the quantile residuals of the LWP is more located inside of the region of the two elliptic curves with few fluctuations. Concerning the graphic qqnorm, the quantile residuals of the LWP are closest to the line $y = x$ compared to other models.

As only the X_{34} was significant for the cure fraction, the estimates of the parameters of the LWP regression model were again estimated with just this covariate. The estimates are presented in Table 9.

Table 9: Estimated values of model parameters final of the LWP regression model to data of cutaneous melanoma.

Parameters	Estimates	S.E.	p-value
α	2.8319	1.5255	-
λ	0.2728	0.0736	-
γ	1.8290	0.1321	
β_0	0.2763	0.1278	0.0306
β_{34}	-1.2143	0.2991	< 0.0001



(a) LEP (b) LEP

(c) LW (d) LW

(e) LWP (f) LWP

Fig. 4: *Worm-Plot* (a; c; e) and qqnorm (b; d; f) of LEP, LW and LWP models for the cutaneous melanoma data.

From the considerations mentioned, the estimated proportion of patients not having cancer recurrence is described by:

$$\hat{p}(\mathbf{x}_i) = \frac{\exp(0.2763 - 1.2143x_{i34})}{1+\exp(0.2763 - 1.2143x_{i34})},\ i = 1,\ldots,417.$$

To detect possible outlying observations, with the support of the software R, we conduct a global influence study to compute the case-deletion measures $GD_i(\theta)$ and $LD_i(\theta)$ presented in Section 6. The influence measures index plots are displayed in Figure 5. From these plots, we noted that the observations 23 and 176 were the most atypical observations concerning the other's observations. Then, we can consider the cases 23 and 176 as possible influential or outlier observations. Thus, these possible atypical observations are patients that have the following characteristics: the observation 23 matches to the patient that has censured and the oldest survival time with the nodule equal 3. The observation 176 matches the patient that has censured and the oldest survival time with the nodule equal 4.

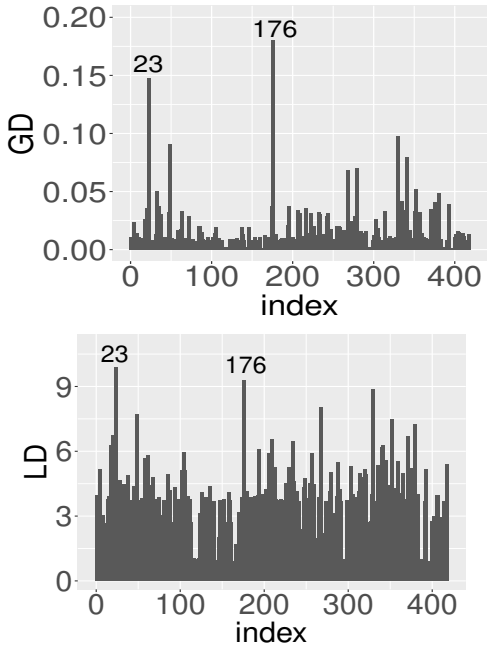


Fig. 5: Index plot for θ: *GD* and *LD* of LWP model for the cutaneous melanoma data.

To verify if these observations are possible influential points, in order to reveal the impact of these observations on the parameter estimates, some combinations of candidates exclusions were made, and the parameters of model were again estimated. Table 10 shows the maximum likelihood estimates and their p-values (among parentheses) these combinations. It can be observed about this table that when the observations 23, and 176 of the data set were removed of form individual or joint form, the estimates remained close concerning the original data, except to the parameters β_0 and β_{34}. Also, the significance of the parameters changes compared to the complete data.

Table 10: Values of the maximum likelihood estimates and p-values of parameters of the LWP regression model.

	Parameters				
Data	α	λ	γ	β_0	β_{34}
A=Complete	2.8319	0.2728	1.8290	0.2763	-1.2143
				(0.0306)	(< 0.001)
A-{23}	2.7724	0.2816	1.8597	0.0794	-0.1204
				(0.5314)	(0.6535)
A-{176}	2.6941	0.2860	1.8636	0.2090	-0.7759
				(0.0997)	(0.0038)
A-{23;176}	2.4092	0.3118	1.9136	0.1449	-0.2966
				(0.2487)	(0.2614)

The Table 11 shows the percentage change of each estimated parameter, that is given by $\left[\left(\widehat{\theta}_j - \widehat{\theta}_{j(i)}\right)/\widehat{\theta}_j\right] X100$, in which $\widehat{\theta}_j$ is the estimate of the maximum likelihood with all observations, and $\widehat{\theta}_{j(i)}$ is the estimate of the maximum likelihood without the i_{th} observation. It can be observed about table 11 that there was an impact of the percentage change when

the observations 23, and 176 were removed of form individual of the data set, in special to the parameters β_0 and β_{34}. Thus, from the analysis made, it was considered the observations 23 and 176 as influence points. However, when theses observations were removed from the individual or joint form, only the significance variable (X_{34}) changed. Then, observations 23 and 176 continued in the analysis.

Table 11: Percentage change of the maximum likelihood estimates of parameters of the LWP regression model.

	Parameters				
Data	α	λ	γ	β_0	β_{34}
A-{23}	2.1010	-3.2316	-1.6788	71.2298	90.0790
A-{176}	4.8643	-4.8332	-1.8877	24.3280	36.1010
A-{23;176}	14.9248	-14.2925	-4.6251	47.546	75.5709

Thus, from the considerations mentioned, we turn to a simplified model that has only the X_{34} as explanatory variable. The estimates for the fitted LWP regression model to the cutaneous melanoma data are listed in Table 9. In this case, the survival function of the LWP regression final model is given by

$$S_{(t|\mathbf{x})} = \frac{\hat{p}(\mathbf{x}_i)\exp(2.8319)}{\exp(2.8319)-1} + \frac{\exp\left\{2.8319\exp\left[-(0.2728t_i)^{1.8290}\right]\right\}}{\exp(2.8319)-1} + \frac{(1-\hat{p}(\mathbf{x}_i))\left(\exp\left\{2.8319\exp\left[-(0.2728t_i)^{1.8290}\right]\right\}\right)}{\exp(2.8319)-1};$$

$i = 1,\ldots,417.$

The final estimated proportion of patients not having cancer recurrence can be described by:

$$\hat{p}(\mathbf{x}_i) = \frac{\exp(0.2763 - 1.2143x_{i34})}{1+\exp(0.2763-1.2143x_{i34})}; i = 1,\ldots,417, \qquad (6)$$

where x_{i34} correspond to the nodules (0=nodule level 2; 1=nodule level 4).

According to the estimated proportion of patients that were not diagnosed the cancer recurrence $(\pi(\mathbf{x}_i))$, the probability of the patients with nodule level 2 not having cancer recurrence is $p^{(X_{34}=0)} = 0.5686 = \hat{p_0}$. The probability of the patients with nodule level 4 not having cancer recurrence is $p^{(X_{34}=1)} = 0.2813 = \hat{p_1}$.

Using the logistic link function, we can make some interpretations about the estimated parameters using the odds ratio. In this case, using the (eq. 6) it is possible to calculate the odds ratio to compare the explanatory variable X_{34} (odds ratio = $\exp(-1.2143) = 0.2969 \cong 30\%$). This value shows that for patients with nodule level 4, the odds of not having cancer recurrence is 70% times smaller patients with nodule level 2

Figure 6 gives the estimated hazard function versus the empirical hazard function as well as the estimated survival function of the final regression. The predicted plot of the final model shows that LWP regression final model fits the data well, hence the estimated curves overlaps the estimates obtained using the Kaplan-Meier estimator and is verified that the estimated hazard function has a unimodal form.

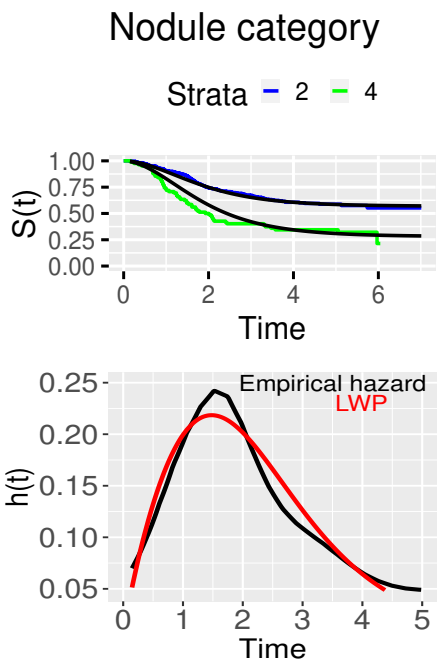


Fig. 6: Survival functions and Fitted hazard using the LWP regression for the cutaneous melanoma data.

9 Concluding Remarks

In this study, the Weibull-Poisson long-term (LWP) regression model was proposed as an alternative to model long-term survival data. This new regression model is based on the Weibull-Poisson distribution in a structure of long-term modeling, in which the covariates were included in the proportion p of cured through the logistic link function and presented as particular case the long-term exponential-Poisson and Weibull regression models.

We used the maximum likelihood method for the estimation of the parameter. Additionally, different simulation studies were adopted to study the means, the biases, and the root of mean squared error of the ML estimates of the proposed model for different values of n and censored observations where it was verified good results.

Finally, an application of the LWP regression model was presented as an alternative for the fit the cutaneous melanoma data. The motivation to study the cutaneous melanoma data is due the disease be responsible for most cases of skin cancer. Through graphical analysis of the TTT-Plot and Kaplan-Meier, the observed values of the AIC criteria and the likelihood ratio test, global influence and residual analysis, it can be noted that the LWP regression model fitted well to the data. Through the proposed model was observed some important characteristics:

- –The probability of the patients with nodule level 2 not having cancer recurrence is 56.86%;
- –The probability of the patients with nodule level 4 not having cancer recurrence is 28.13%;
- –For patients with nodule level 4, the odds of not having cancer recurrence is 70% times smaller patients with nodule level 2;

–The estimated hazard function has a unimodal form.

Thus, it is expected that this model is useful for fitting other datasets.

Acknowledgement

This paper was supported by CNPq and CAPES, Brazil.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

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

[2] V.T. Farewell. The use mixture models for the analysis of survival data with long term survivors, *Biometrics*, **38**, 1041-1046 (1982).

[3] A.Y. Yakovlev, A.D. Tsodikov, and B. Asselain. *Stochastic Models of tumor latency and their biostatistical applications*. Singapore: World Scientifc, 1996.

[4] M. Chen, J. Ibrahim and D. Sinha. A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**, 909-919 (1999).

[5] C.J. Hoggart and J.E. Griffin. A Bayesian partition model for customer attrition. In: George EI (ed.) Bayesian methods with applications to science, policy, and official statistics [selected papers from ISBA 2000]. *Proceedings of the sixth world meeting of the international society for Bayesian analysis.* , Creta, Greece, 28 May 2000-1 June 2000, pp.61-70.

[6] H.M. Chen and J.G Ibrahim Maximum likelihood methods for cure rate models with missing covariates. *Biometrics* **57**, 43-52 (2001).

[7] R. Maller and X. Zhou. Testing for the Presence of Immune or Cured Individuals in Censored Survival Data. *Biometrics*, **51**, 1197-1205 (1995).

[8] J. Haybittle. A two-parameter model for the survival curve of treated cancer patients. *Journal of the American Statistical Association*, **60**, 16-26 (1965)

[9] A. Yakovlev, A. D. Tsodikov, and B. Asselain. *Stochastic models of tumor latency and their biostatistical applications*. Singapore-World Scientific Publishing Company, 288, (1996).

[10] G.S.C. Perdona and F. Louzada-Neto. A general hazard model for lifetime data in the presence of cure rate. *Journal of Applied Statistics*, **38**, 1395-1405 (2011)

[11] A. Perperoglou, A. Keramolpoullos and H.C. Houwelingen. Approaches in modelling long-term survival: An application to breast cancer. *Statistics in Medicine*, **26**, 2666-2685 (2007).

[12] F. Louzada, M. Roman, J.G. Leite and V.G. Cancho. A new long-term survival distribution for cancer data. *Journal of Data Science*, **10**, 241-258 (2012).

[13] V.P. Vigas, J. Mazucheli, F. Louzada. Application of the Weibull-Poisson long-term survival model. *Communications for Statistical Applications and Methods*, **24**, 325-419 (2017).

[14] E.M. Ortega, F.B. Rizzato and C.G. Demétrio, The generalized log-gamma mixture model with covariates: local influence and residual analysis. *Statistical Methods and Applications*, **18**, 305-331 (2009).

[15] N. Kannan, D. Kundu, P. Nair and R. Tripathi. The generalized exponential cure rate model with covariates. *Journal of Applied Statistics*, **10**, 1625-1636 (2010).

[16] M. Castro, V. G. Cancho and J. Rodrigues. A hands-on approach for
fitting long-term survival models under the gamlss framework. *Computer Methods and Programs in Biomedicine*, **97**, 168-177 (2010).

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

[18] Ramires, T. G., Ortega, E. M., Lemonte, A. J., Hens, N., and Cordeiro, G. M. A flexible bimodal model with long-term survivors and different regression structures. *Communications in Statistics-Simulation and Computation*, **49**, 2639-2660 (2018a)

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

[20] Vigas, V. P., Fatoretto, M. B., Slanzon, G. S., Ortega, E. M. M. and Demétrio, C. G. B. and Bittar, C. M. M. Red propolis effect analysis of dairy calves health based on Weibull regression model with long-term survivors. *Research in Veterinary Science*, **136**, 464-471 (2021).

[21] F. Cooner, S. Banerjee, B.P. Carlin and D. Sinha Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association* **102**, 560-572. (2007).

[22] N. Balakrishnan and S. Pal. Log-Normal lifetimes and likelihood-based inference for flexible cure rate models based on com-poisson family. *Computational Statistics and Data Analysis*, **67**, 41-67 (2013).

[23] E.M. Hashimoto, G.M. Cordeiro and E.M. Ortega. The new Neyman type a beta Weibull model with long-term survivors. *Computational Statistics*, **28**, 933-954 (2013).

[24] E.M. Hashimoto, E. M. Ortega, G.M. Cordeiro and V.G. Cancho A new long-term survival model with interval-censored data. *Sankhya B*, **77**, 207-239 (2015).

[25] N. Balakrishnan and S. Pal, (2016). Expectation maximization-based likelihood inference for flexible cure rate models with Weibull lifetimes. *Statistical methods in medical research*, **25**, 1535-1563.

[26] A.K. Suzuki, G.D. Barriga, F. Louzada, V.G. Cancho. A general long-term aging model with different underlying activation mechanisms: Modeling, bayesian estimation, and case in influence diagnostics. *Communications in Statistics-Theory and Methods*, **46**, 3080-3098 (2017).

[27] A.K. Suzuki, V.G. Cancho and F. Louzada. The Poisson-inverse-gaussian regression model with cure rate: a bayesian approach and its case influence diagnostics. *Statistical Papers*, **57**, 133-159 (2016).

[28] E.M. Ortega, G.M. Cordeiro, E.M. Hashimoto, and A.K. Suzuki. Regression models generated by gamma random variables with long-term survivors. *Communications for Statistical Applications and Methods*, **24**, 43-65 (2017).
[29] E.M.P. Bereta, M.A.P. Franco, F. Louzada. The Poisson-Weibull distribution. *Advances and Applications in Statistics*, **22**, 117-118 (2011).
[30] K. Adamidis and S.Loukas, A lifetime distribution with decreasing failure rate. *Statistics Probability Latters*, **39**, 35-42 (1998).
[31] C. Kus. A new lifetime distribution. Computation Statistic. *Data Analysis*, **51**, 4497- 4509 (2007).
[32] M. Chahkandi and M. Ganjali. On some lifetime distributions with decreasing failure rate. *Computational Statistics and Data Analysis*, **53**, 4433-4440 (2009).
[33] W. Barreto-Souza, A.L.D. Morais and G.M. Cordeiro. The Weibull-Geometric Distribution. *Journal of Statistical Computation and Simulation*, **81**, 1-14 (2011).
[34] F. Louzada, M. Roman and V.G. Cancho. The complementary exponential geometric distribution: Model, properties, and a comparison with its counterpart. *Computational Statistics and Data Analysis*, **55**, 2516-2524 (2011).
[35] W. Weibull. A statistical distribution function of wide applicability. *Journal of applied mechanics*, **18**, 293-297 (1951).
[36] R.D. Cook. Detection of influential observations in linear regression. *Technometrics*, **19**, 15-18 (1977).
[37] F. Xie and B. Wei. Diagnostics analysis for log-Birnbaum-Saunders regression models. *Computational Statistics and Data Analysis*, **51**, 4962-4706 (2007).
[38] R.D. Cook and S. Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall (1982).
[39] P. Dunn and G. Smyth. Randomized quantile residuals. *Journal of Computation and Graphical Statistics*, **5**, 236-244 (1996).
[40] S.V. Buuren and M. Fredriks. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, **20**, 1259-1277 (2001).
[41] G. Mudholkar, D. Srivastava and M. Freimer. The exponentiated Weibull family: A reanalysis of the bus-motor-failure data. *Technometrics*, **37**, 436-445 (1995).
[42] G.O. Silva, E.M.M. Ortega, V.G. Cancho and M.L. Barreto. Log-Burr XII regression models with censured Data. *Computational Statistics and Data Analysis*, **52**, 3820-3842 (2008).
[43] G.O. Silva, E.M.M. Ortega, G.M. Cordeiro. log-extended Weibull regression model. *Computational Statistics and Data Analysis*, **53**, 44820-4489 (2009).
[44] V.G. Cancho, F. Louzada-Neto and G.D.C. Barriga. The poisson-exponential lifetime distribution. *Computational Statistics and Data Analysis*, **55**, 677-686 (2011).
[45] V. Leiva, M. Barros, G.A. Paula and M. Galea. Influence diagnostics in log-birnbaum-saunders regression models with censored data. *Computational Statistics and Data Analysis*, **51**, 5694-5707 (2007).
[46] G.O. Silva, E.M. Ortega and G.M. Cordeiro. The beta modified Weibull distribution. *Lifetime data analysis*, **16**, 409-430 (2010).
[47] C.S.M. Paiva, D.M.C. Freire and J.G. Cecatti. Modelos aditivos generalizados para posição, escala e forma (gamlss) na modelagem de curvas de referência. *Revista Brasileira de Ciências da Saúde*, **12**, 289-310 (2010).
[48] G.O. Silva, G.M. Cordeiro and E.M. Ortega. Surviving and non surviving fraction regression models based on the beta modified Weibull distribution. *Model Assisted Statistics and Applications*, **15**, 111-126 (2020).
[49] M.V. Aarset. How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, **15**, 106-108 (1987).

Valdemiro Vigas is a Professor of the Institute of Mathematics at the Federal University of Mato Grosso do Sul, Brazil. He is currently finishing his PhD in Statistic and Agronomic Experimentation from the University of São Paulo, ESALQ-USP, Brazil. His research interests are regression models in survival analysis.

Giovana Oliveira Silva is a PhD in Statistic and Agronomic Experimentation from the University of São Paulo, ESALQ-USP, Brazil (2009). She is currently a Professor at the Department of Statistics at the Federal University of Bahia, Brazil. She works on regression models in survival analysis, Sensitivity analysis, Residual analysis, and new probability distributions.

Josmar Mazucheli is a PhD in Production Engineering (concentration area: Statistics) from the Federal University of Rio de Janeiro (2002). He is currently a Professor at the Department of Statistics at the Maringa State University, Brazil. He works on Reliability/Survival Analysis, Bayesian Inference, Computational Statistics, Biostatistics, and Design of Experiments.