Computational Statistics and Data Analysis 53 (2009) 4482–4489

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A log-extended Weibull regression model

Giovana O. Silva [a], Edwin M.M. Ortega [a,*], Gauss M. Cordeiro [b]

[a] Departamento de Ciências Exatas, ESALQ, Universidade de São Paulo-USP, Piracicaba, Brazil

[b] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco-UFRPE, Recife, Brazil

ARTICLE INFO

Article history:
Received 28 November 2008
Received in revised form 28 April 2009
Accepted 4 July 2009
Available online 10 July 2009

ABSTRACT

A bathtub-shaped failure rate function is very useful in survival analysis and reliability studies. The well-known lifetime distributions do not have this property. For the first time, we propose a location-scale regression model based on the logarithm of an extended Weibull distribution which has the ability to deal with bathtub-shaped failure rate functions. We use the method of maximum likelihood to estimate the model parameters and some inferential procedures are presented. We reanalyze a real data set under the new model and the log-modified Weibull regression model. We perform a model check based on martingale-type residuals and generated envelopes and the statistics *AIC* and *BIC* to select appropriate models.

1. Introduction

The traditional lifetime distributions (Weibull, log-logistic and log-normal) are unable to capture the behavior of a lifetime data that exhibits a bathtub-shaped failure rate curve. These standard distributions are suitable only in situations where the failure rate function is constant, monotone or unimodal. However, this function may frequently present a bathtub-shaped form. The distributions that exhibit bathtub-shaped failure rate are usually complex and, therefore, difficult to be modeled. Thus, it is important to propose new distributions that do present much practicability to model this type of failure rate.

Some distributions have been introduced to model bathtub-shaped data, such as the generalized gamma distribution (Stacy, 1962), the generalized *F* distribution (Prentice, 1974), the IDB distribution (Hjorth, 1980), the exponential-power family (Smith and Bain, 1975), among others. A good review of these models is presented in Rajarshi and Rajarshi (1988). In the last decade, new classes of distributions have been proposed based on extended forms of the Weibull distribution for modeling data of this kind from a desire to provide a better fitting than the Weibull distribution. Most of the generalizations or modifications of the Weibull distribution that appeared since 2004 have been discussed by Pham and Lai (2007).

Although many distributions are discussed in the literature to accommodate the bathtub-shaped failure rate, a few regression models have been proposed with this objective, among them, the log-exponentiated-Weibull (Cancho et al., 1999), generalized log-F (Kalbfleisch and Prentice, 2002), generalized log-gamma (Lawless, 2003) and log-modified Weibull (Carrasco et al., 2008) regression models.

Regression models can be proposed in different forms in survival analysis. For example, the location-scale regression model (Klein and Moeschberger, 1997; Lawless, 2003) is distinguished and it is frequently used in clinical trials. In this paper, we propose a new regression model using the logarithm of the extended Weibull distribution (Xie et al., 2002). The modification of the existing distribution leads to a location-scale regression model suitable for fitting censored survival times

* Corresponding address: Departamento de Ciências Exatas, USP, Av. Pádua Dias 11, Caixa Postal 9, 13418-900 Piracicaba, São Paulo, Brazil. Tel.: +55 11 19 34294127; fax: +55 11 19 34294468.

E-mail addresses: giovana@ufba.br (G.O. Silva), edwin@esalq.usp.br (E.M.M. Ortega), gausscordeiro@uol.com.br (G.M. Cordeiro).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.07.004

with bathtub-shaped hazard rates referred to as the log-extended Weibull (LEW) regression model. We can also check the distributional assumptions of the model by examining the residuals (see, for example, Barlow and Prentice (1988), Therneau et al. (1990) and Collett (2003)).

In Section 2, we define the LEW distribution and derive its moments. In Section 3, we propose a LEW regression model of location-scale form and obtain the maximum likelihood estimates. We also review a generalized likelihood ratio test that can be used for comparing nonnested models. In Section 4, we provide expressions of martingale-type residuals for the LEW regression model. We show in Section 5 that the proposed model is more adequate to fit the lung cancer data analysis than log-modified Weibull (LMW) regression model proposed by Carrasco et al. (2008), by checking the residual plots for both models and discriminating between the models using three different statistics. Section 6 ends with some conclusions.

2. A log-extended Weibull distribution

Most generalized Weibull distributions have been proposed in reliability literature to provide a better fitting of certain data sets than the traditional two- or three-parameter Weibull model. See, for example, the distributions listed and discussed in Tables I and II given by Pham and Lai (2007). A very complicated generalized Weibull distribution often diminishes the probability of interpreting the parameters and a generalization that has more than three parameters is undesirable. Xie et al. (2002) introduced a three-parameter Weibull distribution, the so-called the extended Weibull distribution, with the probability density function (pdf) defined by

$$f(t;\lambda,\tau,\alpha)=\lambda\tau\left(\frac{t}{\alpha}\right)^{\tau-1}\exp\left\{\left(\frac{t}{\alpha}\right)^{\tau}+\lambda\alpha\left[1-\exp\left(\left(\frac{t}{\alpha}\right)^{\tau}\right)\right]\right\},\quad t\geq 0, \tag{1}$$

where $\lambda>0$ and $\alpha>0$ are scale parameters and $\tau>0$ is a shape parameter. The corresponding survival and failure rate functions are, respectively, given by

$$S(t;\lambda,\tau,\alpha)=\exp\left\{\lambda\alpha\left[1-\exp\left(\left(\frac{t}{\alpha}\right)^{\tau}\right)\right]\right\}\quad\text{and}\quad h(t;\lambda,\tau,\alpha)=\lambda\tau\left(\frac{t}{\alpha}\right)^{\tau-1}\exp\left[\left(\frac{t}{\alpha}\right)^{\tau}\right].$$

The failure rate function of the extended Weibull distribution has a bathtub shape when $\tau<1$ and an increasing function when $\tau\geq 1$ (Xie et al., 2002). This distribution is mainly related to the model studied by Chen (2000) with an additional scale parameter α. When $\alpha\to\infty$, $1-\exp[(t/\alpha)^{\tau}]\approx-(t/\alpha)^{\tau}$, so that the survival function converges to the limit $S(t;\lambda,\tau,\alpha)\approx\exp\{-\lambda\alpha^{1-\tau}t^{\tau}\}$, which is the Weibull distribution with shape parameter τ and scale parameter $\alpha^{\tau-1}\lambda^{-1}$. Hence, the extended Weibull distribution has the Weibull distribution as a special and asymptotic case.

The extended Weibull distribution is easily simulated using the inverse probability method. If U is a uniform random variable on the interval $(0,1)$, then the random variable defined by $T=\{\alpha^{\tau}\log[1-(\lambda\alpha)^{-1}\log(1-U)]\}^{\frac{1}{\tau}}$ follows the extended Weibull distribution (1). The distribution of the logarithm $Y=\log(T)$ of the random variable T is called the LEW distribution, parameterized in terms of the parameters $\sigma=\tau^{-1}$, $\mu=\log(\alpha)$ and λ, and its pdf has the form

$$f(y;\lambda,\sigma,\mu)=\frac{\lambda}{\sigma}\exp\left(\frac{y-\mu}{\sigma}\right)\exp\left\{\mu+\exp\left(\frac{y-\mu}{\sigma}\right)+\lambda\exp(\mu)\left[1-\exp\left[\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right]\right\}, \tag{2}$$

where $-\infty<y<\infty$, $\lambda>0$, $\sigma>0$ and $-\infty<\mu<\infty$. The corresponding survival function reduces to

$$S(y;\lambda,\sigma,\mu)=\exp\left\{\lambda\,\exp(\mu)\left[1-\exp\left[\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right]\right\}.$$

Further, after suitable transformation, we define the standard random variable $Z=(Y-\mu)/\sigma$ with density function

$$f(z;\lambda,\mu)=\lambda\exp\{z+\mu+\exp(z)+\lambda\exp(\mu)[1-\exp[\exp(z)]]\},\quad -\infty<z<\infty. \tag{3}$$

Plots of the density function (3) for selected parameter values are shown in Fig. 1. Eq. (3) for the standardized LEW distribution will be used in Section 3 to specify the error distribution of an accelerated failure time model.

The sth moment of the extended Weibull density (1) when $s/\tau=m$ is an integer was recently obtained by Nadarajah (2005) as

$$\mu'_s=m\alpha^s\exp(\lambda\alpha)\left.\frac{\partial^{m-1}(\lambda\alpha)^{-\nu}\gamma(\nu,\lambda\alpha)}{\partial\nu^{m-1}}\right|_{\nu=0}, \tag{4}$$

where $\gamma(\nu,\lambda\alpha)=\int_0^{\lambda\alpha}w^{\nu-1}e^{-w}dw$ is the well-known incomplete gamma function. Some special cases of Eq. (4) for $\tau=1/2$ and $1/3$ and $s=1$ and 2 are given by Nadarajah (2005). By expanding $Y^s=\log(T)^s$ in Taylor series around μ'_1, the sth moment of Y can be written as

$$E(Y^s)=\log(\mu'_1)^s+\sum_{i=2}^{\infty}\frac{G^{(i)}(\mu'_1)\,\mu_i}{i!},$$

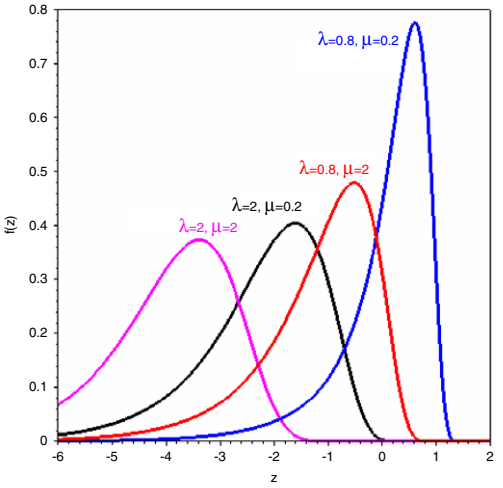

Fig. 1. Plots of the density (3) for $\lambda = 0.8$ or 2 and $\mu = 0.2$ or 2.

where $G^{(i)}(\mu_1')$ is the ith derivative of $G(\mu_1') = \log(\mu_1')^s$ with respect to μ_1' and $\mu_i = E(T - \mu_1')^i$ is the ith central moment of T. Expressing the central moments of T in terms of their ordinary moments, the sth ordinary moment of Y can be written as weighted infinite sums of products of suitable ordinary moments of T by powers of the expected value μ_1', namely

$$E(Y^s) = \log(\mu_1')^r + \sum_{i=2}^{\infty}\sum_{k=0}^{i}\frac{(-1)^k G^{(i)}(\mu_1')\mu_{i-k}'\mu_1'^k}{(i-k)!k!}, \tag{5}$$

where the moments μ_{i-k}' and μ_1' are readily obtained from Eq. (4). Formula (5) holds for any sth moment only when s/τ is an integer. The derivatives of $G(\mu_1') = \log(\mu_1')^s$ are easily calculated in Maple up to any order. For example, we obtain

$$G^{(4)}(\mu_1') = s\{(s^3 - 6s^2 + 11s - 6)\delta^{s-4} - 6(s^2 - 3s + 2)\delta^{s-3} + 11(s-1)\delta^{s-2} - 6\delta^{s-1}\}/\mu_1'^4,$$

where $\delta = \log(\mu_1')$.

Hence, the ordinary moments of the LEW distribution is a function of the parameters μ, σ and λ. Clearly, the moments of Z can be easily obtained from the moments of Y.

3. A log-extended Weibull regression model

In many practical applications, the lifetimes are affected by explanatory variables such as the cholesterol level, blood pressure, weight and many others. Parametric models for estimating univariate survival functions and for the censored data regression problems are widely used. When the parametric models provide a good fit to the lifetime data set, they tend to give more precise estimates of the quantities of interest because these estimates are based on fewer parameters. Let $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ be the explanatory variable vector associated with the ith response variable $y_i = \log(t_i)$, note that, y_i is the logarithm of the survival time t_i. Based on the LEW distribution, a linear regression model linking the response variable y_i and the explanatory variable vector $\mathbf{x}_i$ can be defined by

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \sigma z_i, \quad i = 1, \ldots, n, \tag{6}$$

where the random error z_i follows the density function (3), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, $\sigma > 0$ and $\lambda > 0$ are unknown parameters and $\mathbf{x}_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{ip})$ is the explanatory variable vector modeling the linear predictor $\mu_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. Hence, the linear predictor vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ of the LEW regression model is simply $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ is a known model matrix. The log-Weibull (or extreme value) regression model is obtained as a special case from (6) when $\alpha \to \infty$.

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 and 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. The total log-likelihood function for the model parameters $\boldsymbol{\theta} = (\lambda, \sigma, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ follows from (3) and (6) as

$$\begin{aligned} l(\boldsymbol{\theta}) &= r\log(\lambda) - r\log(\sigma) + \sum_{i\epsilon F}\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \sum_{i\epsilon F} z_i + \sum_{i\epsilon F}\exp(z_i) \\ &\quad + \sum_{i\epsilon F}\lambda\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\{1 - \exp[\exp(z_i)]\} + \sum_{i\epsilon C}\lambda\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\{1 - \exp[\exp(z_i)]\}, \end{aligned} \tag{7}$$

where r is the number of uncensored observations (failures) and $z_i = (y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})/\sigma$. The maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\theta}}$ of the parameter vector $\boldsymbol{\theta} = (\lambda, \sigma, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ of the LEW regression model can be obtained by maximizing the log-likelihood function (7). The estimation process is straightforward and we use the matrix programming language Ox (MAXBFGS subroutine) (see Doornik (2007)) to compute the estimate $\widehat{\boldsymbol{\theta}}$.

After fitting the model (6), the survival function for Y (the survival function for T comes easily by inverting the equation), say $P(Y \geq y) = S(y; \lambda, \sigma, \boldsymbol{\beta}^{\mathrm{T}})$, can be readily estimated by

$$S(y; \widehat{\lambda}, \widehat{\sigma}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}}) = \exp\left\{\widehat{\lambda}\exp(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\left[1 - \exp\left[\exp\left(\frac{y - \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{\widehat{\sigma}}\right)\right]\right]\right\}.$$

Under conditions that are fulfilled for the parameter vector $\boldsymbol{\theta}$ in the interior of the parameter space but not on the boundary, the asymptotic distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is multivariate normal $N_{p+2}(0, K(\boldsymbol{\theta})^{-1})$, 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 $(p+1)\times(p+1)$ observed information matrix $-\ddot{\mathbf{L}}_{\theta\theta}$ and then the asymptotic inference for the parameter vector $\boldsymbol{\theta}$ can be based on the normal approximation $N_{p+2}(0, -\ddot{\mathbf{L}}_{\theta\theta}^{-1})$ for $\widehat{\boldsymbol{\theta}}$. The elements of the observed information matrix

$$-\ddot{\mathbf{L}}_{\theta\theta} = \{-\ddot{L}_{r,s}\} = \begin{pmatrix} -\mathbf{L}_{\lambda\lambda} & -\mathbf{L}_{\lambda\sigma} & -\mathbf{L}_{\lambda\beta_j} \\ . & -\mathbf{L}_{\sigma\sigma} & -\mathbf{L}_{\sigma\beta_j} \\ . & . & -\mathbf{L}_{\beta_j\beta_s} \end{pmatrix}$$

for $j, s = 1, \ldots, p$ are given in Appendix.

The asymptotic multivariate normal $N_{p+2}(0, -\ddot{\mathbf{L}}_{\theta\theta}^{-1})$ distribution can be used to construct approximate confidence intervals for some parameters in $\boldsymbol{\theta}$ and for the hazard and survival functions. In fact, an $100(1-\gamma)\%$ asymptotic confidence interval for each parameter θ_r can be expressed as

$$ACI_r = \left(\hat{\theta}_r - z_{\gamma/2}\sqrt{-\widehat{\ddot{L}}^{r,r}}, \hat{\theta}_r + z_{\gamma/2}\sqrt{-\widehat{\ddot{L}}^{r,r}}\right),$$

where $-\widehat{\ddot{L}}^{r,r}$ denotes the rth diagonal element of the inverse of the estimated observed information matrix $-\widehat{\ddot{\mathbf{L}}}_{\theta\theta}^{-1}$ and $z_{\gamma/2}$ is the quantile $1-\gamma/2$ of the standard normal distribution. The asymptotic normality is also useful for testing goodness of fit of some sub-models and for comparing some special sub-models using the likelihood ratio (LR) statistic.

We consider the partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\theta}_1$ is a subset of the parameters of interest and $\boldsymbol{\theta}_2$ is a subset of the remaining parameters. The LR statistic for testing the null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^{(0)}$ versus the alternative hypothesis $H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_1^{(0)}$ is given by $w = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widetilde{\boldsymbol{\theta}})\}$, where $\widetilde{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}$ are the estimates under the null and alternative hypotheses, respectively. The statistic w is asymptotically (as $n \to \infty$) distributed as χ_k^2, where k is the dimension of the subset of parameters $\boldsymbol{\theta}_1$ of interest.

3.1. Discriminating among nonnested models

Carrasco et al. (2008) introduced a location-scale regression model based on the modified Weibull distribution that has a bathtub-shaped failure rate function. Thus, the LEW regression model (6) seems a good alternative to their model. However, these two models are nonnested. For comparison of nonnested survival models, Klein and Moeschberger (1997) suggested the criterions *AIC* (Akaike information criterion) and *BIC* (Bayesian information criterion) given by

$$AIC = -2\log(likelihood) + 2(p + 2 + k) \quad \text{and} \quad BIC = -2\log(likelihood) + (p + k)\log(n),$$

where p is the number of estimated parameters and $k = 2$ for both models. The model with the smallest criterion (AIC or BIC value) can be selected as the preferred model.

An alternative generalized LR statistic which can be used for discriminating among nonnested models is discussed in the book of Cameron and Trivedi (1998, p. 184). Consider choosing between two nonnested models - model F_θ with density function $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ and model G_γ with density function $g(y_i|\mathbf{x}_i, \boldsymbol{\gamma})$. This statistic is a distance between the two models measured in terms of the Kullback–Liebler information criterion. It is defined by

$$T_{LR,NN} = \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\log\frac{f(y_i|\mathbf{x}_i, \widehat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i, \widehat{\boldsymbol{\gamma}})}\right\} \div \left\{\frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{f(y_i|\mathbf{x}_i, \widehat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i, \widehat{\boldsymbol{\gamma}})}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(y_i|\mathbf{x}_i, \widehat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i, \widehat{\boldsymbol{\gamma}})}\right)^2\right\}. \tag{8}$$

For strictly nonnested models, the statistic (8) converges in distribution to a standard normal distribution under the null hypothesis of equivalence of the models. Thus, the null hypothesis is not rejected if $|T_{LR,NN}| \leq z_{\frac{\alpha}{2}}$. On the other hand, we reject at significance level α the null hypothesis of equivalence of the models in favor of model F_θ being better (or worse) than model G_γ if $T_{LR,NN} > z_\alpha$ (or $T_{LR,NN} > -z_\alpha$).

We shall use (8) in Section 5 for comparing the fitted LEW and LMW models.

4. Residual analysis

As is already known, departures from the error assumption as well as the presence of outliers in regression models involve examining the residuals. The martingale residuals are recommended in counting processes and they are defined by $r_{M_i} = \delta_i + \log\{S(y_i; \hat{\boldsymbol{\theta}})\}$, where $\delta_i = 0$ ($\delta_i = 1$) denotes a censored (uncensored) observation and $S(y_i; \widehat{\boldsymbol{\theta}})$ was presented in Section 3. A disadvantage of the martingale residual is that the distribution of r_{M_i} is markedly skewed, and so it fails to have similar properties to those of the normal distribution. In fact, r_{M_i} ranges from a minimum value $-\infty$ to a maximum value $+1$. For lifetime regression models some extensions are needed to the above definition for model checking procedures. A transformation to make the distribution of the transformed residual as normal as possible would be more appropriate for performing residual analysis.

Therneau et al. (1990) discussed a possible transformation of the martingale residual based on the deviance component residual for Cox's proportional hazard model with no time-dependent explanatory variables. It turns out that the ith martingale-type residual can be written as

$$r_{D_i} = \text{sign}(r_{M_i})\{-2[r_{M_i} + \delta_i \log(\delta_i - r_{M_i})]\}^{\frac{1}{2}},$$

where r_{M_i} is the corresponding martingale residual. A motivation for this transformation is to obtain a new residual symmetrically distributed around zero. A more extensive examination of this residual is given by Leiva et al. (2007) and Ortega et al. (2008). Hence, the martingale-type residual for the LEW regression model is equal to

$$r_{D_i} = \begin{cases} \sqrt{2}\,\text{sign}\left\{1 + \hat{\lambda}\exp(\hat{\mu})\left[1 - \exp\{\exp(\hat{z}_i)\}\right]\right\}\left\{-1 - \hat{\lambda}\exp(\hat{\mu})\left[1 - \exp(\exp(\hat{z}_i))\right]\right. \\ \quad \left. - \log\left\{-\hat{\lambda}\exp(\hat{\mu})\left[1 - \exp(\exp(\hat{z}_i))\right]\right\}\right\}^{\frac{1}{2}} & \text{if } i \in F \\ \sqrt{2}\,\text{sign}\left\{\hat{\lambda}\exp(\hat{\mu})\left[1 - \exp(\exp(\hat{z}_i))\right]\right\}\left\{-\hat{\lambda}\exp(\hat{\mu})\left[1 - \exp(\exp(\hat{z}_i))\right]\right\}^{\frac{1}{2}} & \text{if } i \in C. \end{cases}$$

We have developed some Monte Carlo simulations for the LEW regression model that indicate that the empirical distribution of the martingale-type residual is in agreement with the standard normal distribution. Further, Ortega et al. (2008) showed that the same result holds for the log-Weibull regression model. We can use normal probability plots for r_{D_i} with simulated envelopes for both LEW and LMW models, as suggested by Atkinson (1985), obtained as follows: (i) fit the model and generate a sample of n-independent observations using the fitted model as if it was the true model; (ii) fit the model to the generated sample using $(\delta_i, \mathbf{x}_i)$ as the data set and compute the values of the residuals; (iii) repeat steps (i) and (ii) m times; (iv) obtain ordered values of the residuals, $r^*_{(i)v}$, $i = 1, 2, \ldots, n$ and $v = 1, 2, \ldots, m$; (v) consider n sets of m ordered statistics and for each set compute the mean, minimum and maximum values; (vi) plot these values and the ordered residuals of the original sample against the normal scores. The minimum and maximum values of the m ordered statistics yield the envelope. Observations corresponding to residuals outside the limits provided by the simulated envelope need further investigation. Additionally, if a considerable proportion of points falls outside the envelope, then we have evidence against the adequacy of the fitted model. Plots of such residuals against the fitted values can also be useful.

5. Lung cancer survival data

In order to demonstrate the proposed methodology, we use the lung cancer data set reported by Prentice (1973) referring to the survival time (t) and the explanatory variables: performance status at diagnosis (x_1), a measure of general fitness on a scale from 0 to 100, the age of the patient (x_2) and the number of months from diagnosis of cancer (x_3). In addition, each patient was assigned one of two chemotherapy treatments (standard or test) and the tumors were classified into four types: large, adeno, small and squamous. The data contain $n = 40$ observations of which 3 are censored. Lawless (2003) fitted a Weibull regression model to analyze these data which was reasonable as a first model.

Initially, we consider a device called the total time on test (TTT) plot (Aarset, 1987), which can help us in choosing a particular model. The TTT plot is obtained by plotting $\mathrm{G}(r/n) = [\sum_{i=1}^{r} T_{i:n} + (n-r)T_{r:n}]/\sum_{i=1}^{n} T_{i:n}$, where $r = 1, \ldots, n$ and $T_{i:n}$, for $i = 1, \ldots, n$, are the order statistics of the sample, against r/n (Mudholkar et al., 1996). Fig. 2 shows the TTT plot for these data.

We centered only the explanatory variables x_1, x_2 and x_3 (Lawless, 2003) and work with the following model

$$y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \beta_3(x_{i3} - \bar{x}_3) + \sum_{p=4}^{7} \beta_p x_{ip} + \sigma z_i, \quad i = 1, \ldots, 40, \tag{9}$$

where variable $y_i = \log(t_i)$ follows the LEW distribution given in (2), the random error z_i is specified by the standard LEW distribution (3) and

$x_{i4} = 1$ if tumor type is squamous, 0 otherwise;
$x_{i5} = 1$ if tumor type is small, 0 otherwise;
$x_{i6} = 1$ if tumor type is adeno, 0 otherwise;
$x_{i7} = 0$ if treatment is test, 1 otherwise.

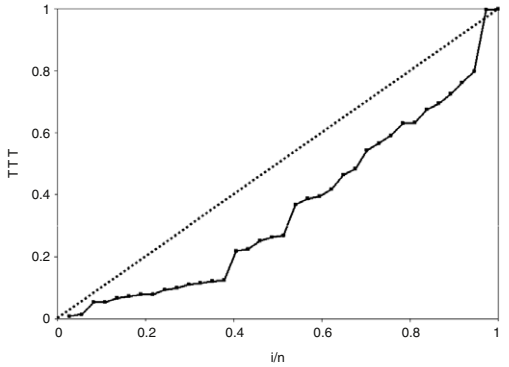

Fig. 2. TTT plot on lung cancer survival data.

Table 1
Estimates of the parameters for the LEW and LMW regression models fitted to the complete lung cancer data set.

Parameter	LEW			LMW		
	Estimate	SE	p-value	Estimate	SE	p-value
λ	0.0020	0.0007	–	–	–	–
α	–	–	–	−4.5636	0.7811	–
σ	1.3116	0.2772	–	1.2930	0.2676	–
β_0	4.3363	0.8442	<0.0001	5.4027	0.5656	<0.0001
β_1	0.0647	0.0065	<0.0001	0.0673	0.0090	<0.0001
β_2	0.0433	0.0180	0.0161	0.0331	0.0184	0.0727
β_3	0.0972	0.0210	<0.0001	0.0830	0.0224	0.0002
β_4	3.4403	2.0213	0.0888	2.6831	1.9948	0.1786
β_5	−0.5825	0.2973	0.0501	−0.5324	0.3360	0.1131
β_6	−2.3091	0.5074	<0.0001	−2.0155	0.5457	0.0002
β_7	−0.4672	0.3213	0.1459	−0.3640	0.3571	0.3080

We fitted the LMW and LEW regression models to these data. Table 1 gives the estimates (and their standard errors) of the parameters for both regression models. Then, we select the best model based on the values of the statistics *AIC* and *BIC*. The statistic *AIC* yields the value 130.248 for the LMW model and 128.208 for the LEW model, whereas the statistic *BIC* yields 139.551 for the LMW model and 137.51 for the LEW model. The values of these statistics indicate that the LEW regression model is more adequate to explain the data set than the LMW model. Additionally, we perform the LR test of nonnested models as described in Section 3.1, where here $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ and $g(y_i|\mathbf{x}_i, \boldsymbol{\gamma})$ denote the density function (2) of our model and the density (5) of Carrasco et al.'s (2008) model, respectively. The generalized LR test statistic yields $TR_{LR,NN} = 16.1520$. Since $TR_{LR,NN} > 1.96$, we reject at significance level 0.05 the null hypothesis of equivalence of the LMW and LEW models. Further, the value of this statistic is in agreement with the previous result and really help us in selecting the LEW regression model.

The current estimates of the regression parameters are similar for the LEW and LMW models but their standard errors are different. Thus, the conclusions may be different for both models. We continue the analysis through the residual plots, which are useful to evaluate both fitted models. In order to detect possible outliers and departures from the error distributional assumptions of the LMW and LEW models, Figs. 3 and 4 show the plots of the residuals against the fitted values and the normal plots, where both generated envelopes are calculated for the martingale-type residuals.

Fig. 3(a) indicates that the residuals are not randomly scattered around zero for the LMW model. This plot also shows that the residuals of the observations 10 and 25 are possible outliers, i.e. are not in the interval $(-3, 3)$. These observations are uncensored and have smaller survival times. The appearance of Fig. 3(b) gives a much better randomly scattered plot of the residuals around zero for the LEW model. It also shows that the LEW regression model is more appropriate to fit the data since it does not present outliers.

Further, the envelope plots in Fig. 4(a) and (b) of the martingale-type residuals against the order statistics of the normal distribution for both models clearly indicate that the LEW distribution is more suitable for modeling the current data than the LEW distribution.

In summary, we recommend using the LEW regression model based on the above analysis. Table 1 suggests that x_1, x_2 and x_3 are significant and we can interpret the estimated coefficients as follows: the expected survival time should increase approximately 4%($e^{0.0433} \times 100\%$) as the center age (x_2) increases one unit, the other variables being fixed. Similar analysis could be done for the variables x_1 and x_3. In addition, the treatment does not appear to have sizeable effects, but the adeno tumor type is an important feature. As the estimate of β_6 is negative, the patients whose tumor type is adeno present smaller survival probabilities than those patients with large tumor types.

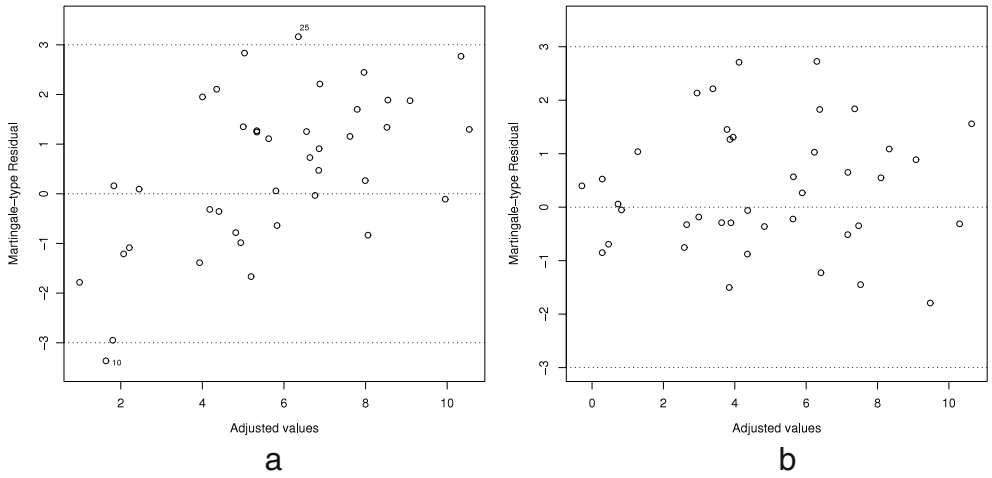

Fig. 3. Plots of the Martingale-type residuals against the adjusted values from the (a) LMW regression model and (b) LEW regression model.

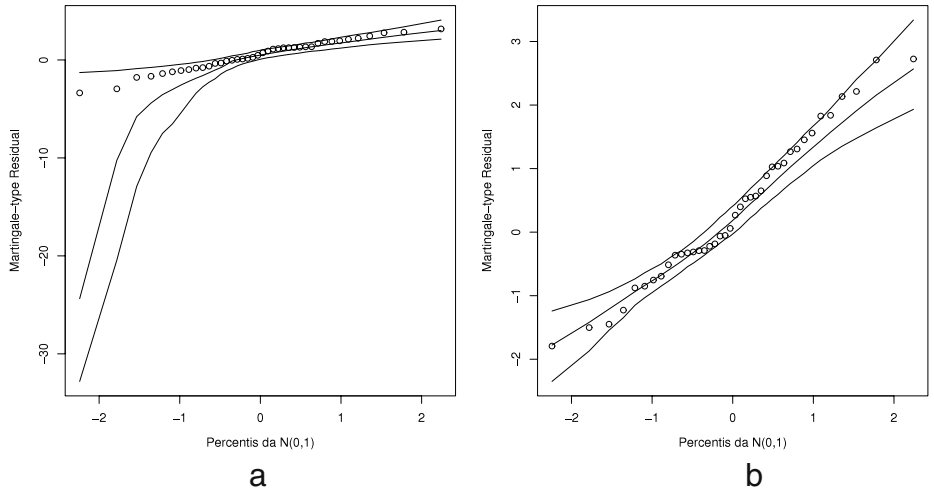

Fig. 4. Plots of the martingale-type residuals against the order statistics of the normal distribution from the (a) LMW regression model and (b) LEW regression model.

6. Concluding remarks

There is an extensive literature on the Weibull distribution for modeling lifetime data. However, the Weibull distribution does not exhibit a bathtub-shaped failure rate function and thus it cannot be used to model several lifetime data sets. To cope with this situation, several generalizations or modifications of the Weibull distribution have been published recently (see, Pham and Lai (2007)). Xie et al. (2002) introduced a modified Weibull distribution which exhibits bathtub-shaped failure rate functions. We define a new distribution via the logarithm of the modified Weibull distribution, the so-called log-extended Weibull (LEW) distribution, which is able to capture the behavior of a lifetime data set that has a bathtub-shaped failure rate function. Further, based on this new distribution, we develop a LEW regression model to be competitive to the log-modified Weibull (LMW) regression model proposed by Carrasco et al. (2008). A lung cancer real data set is reanalyzed to show the performance of the proposed regression model. In fact, we show that the LEW regression model has better performance than the LMW regression model for these data. The codes of the programs used for fitting the LEW regression model are available from the authors upon request.

Acknowledgments

This work was supported by CNPq and CAPES. The authors are grateful to two anonymous referees and the Editor for very useful comments and suggestions.

Appendix. Matrix of second derivatives $\ddot{\mathbf{L}}_{\theta\theta}$

We give the necessary formulas to obtain the second-order partial derivatives of the log-likelihood function. After some algebraic manipulations, we obtain

$$\mathbf{L}_{\lambda\lambda} = -\frac{r}{\lambda^2},$$

$$\mathbf{L}_{\lambda\sigma} = \frac{1}{\sigma}\sum_{i=1}^{n} z_i h_i \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}),$$

$$\mathbf{L}_{\lambda\beta_j} = \sum_{i=1}^{n} x_{ij}\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \frac{1}{\sigma}\sum_{i=1}^{n} x_{ij}\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})h_i\left[1-\sigma\exp(-z_i)\right],$$

$$\mathbf{L}_{\sigma\sigma} = \frac{r}{\sigma^2} + \frac{1}{\sigma^2}\sum_{i\epsilon F}\{z_i\left[2+\exp(z_i)(z_i+2)\right]\} + \frac{\lambda}{\sigma^2}\sum_{i=1}^{n} z_i\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})h_i\left[-2-z_i(1+\exp(z_i))\right],$$

$$\mathbf{L}_{\sigma\beta_j} = \frac{1}{\sigma^2}\sum_{i\epsilon F} x_{ij}\left[1+\exp(z_i)(1+z_i)\right] + \frac{\lambda}{\sigma^2}\sum_{i=1}^{n} x_{ij}\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})h_i\left[-1+z_i(\sigma-1-\exp(z_i))\right],$$

$$\mathbf{L}_{\beta_j\beta_s} = \frac{1}{\sigma^2}\sum_{i\epsilon F} x_{ij}x_{is}\exp(z_i) + \lambda\sum_{i=1}^{n} x_{ij}x_{is}\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \frac{\lambda}{\sigma^2}\sum_{i=1}^{n} x_{ij}x_{is}\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})h_i\left[2\sigma-1-\sigma^2\exp(-z_i)-\exp(z_i)\right],$$

for $j, s = 1, 2, \ldots, p$, where $h_i = \exp\{z_i + \exp(z_i)\}$ and $z_i = \frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma}$.

References

Aarset, M.V., 1987. How to identify bathtub hazard rate. IEEE Transactions on Reliability 36, 106–108.

Atkinson, A.C., 1985. Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostics Regression Analysis, 2nd ed. Clarendon Press, Oxford.

Barlow, W.E., Prentice, R.L., 1988. Residual for relative risk regression. Biometrika 75, 65–74.

Cancho, V., Bolfarine, H., Achcar, J.A., 1999. A Bayesian analysis for the exponentiated-Weibull distribution. Journal of Applied Statistical Science 8, 227–242.

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.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, New York.

Chen, Z.A., 2000. A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. Statistics and Probability Letters 49, 155–161.

Collett, D., 2003. Modelling Survival Data in Medical Research, 2nd ed. Chapman and Hall, London.

Doornik, J., 2007. Object-oriented Matrix Programming using Ox, 5th ed. Timberlake Consultants Ltd., London.

Hjorth, U., 1980. A reliability distribution with increasing, decreasing, constant and bathtub failure rates. Technometrics 22, 99–107.

Kalbfleisch, J.D., Prentice, R.L., 2002. The Statistical analysis of Failure Time Data, 2nd ed. John Wiley, New York.

Klein, J.P., Moeschberger, M.L., 1997. Survival Analysis: Techniques for Censored and Truncated Data. Springer, New York.

Lawless, J.F., 2003. Statistical Models and Methods for Lifetime Data, 2nd ed. John Wiley, New York.

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.

Mudholkar, G.S, Srivastava, D.K., Friemer, M., 1996. A generalization of the Weibull distribution with application to the analysis of survival data. Journal of the American Statistical Association 91, 1575–1583.

Nadarajah, S., 2005. On the moments of the modified Weibull distribution. Reliability Engineering and System Safety 90, 114–117.

Ortega, E.M.M., Paula, G.A., Bolfarine, H., 2008. Deviance residuals in generalized log-gamma regression models with censored observations. Journal of Statistical Computation and Simulation 78, 747–764.

Pham, H., Lai, Chin-Diew, 2007. On recent generalizations of the Weibull distribution. IEEE Transactions on Reliability 56, 454–458.

Prentice, R.L., 1973. Exponential survival with censoring and explanatory variables. Biometrika 60, 279–288.

Prentice, R.L., 1974. A log-gamma model and its maximum likelihood estimation. Biometrika 61, 539–544.

Rajarshi, S., Rajarshi, M.B., 1988. Bathtub distributions: A review. Communication in Statistics - Theory and Methods 17, 2597–2621.

Smith, R.M., Bain, L.J., 1975. An exponential power life testing distributions. Communications in Statistics 4, 469–481.

Stacy, E.W., 1962. A generalization of the gamma distribution. Annals of Mathematical Statistics 33, 1187–1192.

Therneau, T.M., Grambsch, P.M, Fleming, T.R., 1990. Martingale-based residuals for survival models. Biometrika 77, 147–160.

Xie, M., Tang, Y., Goh, T.N., 2002. A modified Weibull extension with bathtub failure rate function. Reliability Engineering and System Safety 76, 279–285.