Birnbaum–Saunders frailty regression models: Diagnostics and application to medical data

Jeremias Leão[1,2], **Víctor Leiva***[,3,4], **Helton Saulo**[5,6], and **Vera Tomazella**[2]

[1] Department of Statistics, Universidade Federal do Amazonas, Manaus, Brazil
[2] Department of Statistics, Universidade Federal de São Carlos, São Carlos, Brazil
[3] Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Viña del Mar, Chile
[4] School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
[5] Institute of Mathematics and Statistics, Universidade Federal de Goiás, Goiânia, Brazil
[6] Department of Statistics, Universidade de Brasília, Brasília, Brazil

Received 13 January 2016; revised 9 October 2016; accepted 21 October 2016

In survival models, some covariates affecting the lifetime could not be observed or measured. These covariates may correspond to environmental or genetic factors and be considered as a random effect related to a frailty of the individuals explaining their survival times. We propose a methodology based on a Birnbaum–Saunders frailty regression model, which can be applied to censored or uncensored data. Maximum-likelihood methods are used to estimate the model parameters and to derive local influence techniques. Diagnostic tools are important in regression to detect anomalies, as departures from error assumptions and presence of outliers and influential cases. Normal curvatures for local influence under different perturbations are computed and two types of residuals are introduced. Two examples with uncensored and censored real-world data illustrate the proposed methodology. Comparison with classical frailty models is carried out in these examples, which shows the superiority of the proposed model.

Keywords: Birnbaum–Saunders distribution; Censored data; Global and local influence; Maximum-likelihood method; Residual analysis.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

A common assumption in survival models is to suppose that the population of patients under study is homogeneous and with the same structure for its covariates. However, patients who apparently have similar observable attributes, such as age, gender, or weight, may relapse or die at distinct instants of time due to, for example, environmental and/or genetic factors. Therefore, unobservable random effects could be incorporated to survival models for improving its accuracy. Hereafter, we refer to relapses or death times of patients as lifetimes.

Vaupel et al. (1979) proposed models that consider an unobservable random effect called frailty. The frailty indicates that apparently similar patients can have different risks. Thus, different patients with similar values for their observed covariates may possess distinct frailties. Then, frailer patients tend to experience the event earlier than those who are less frail. Therefore, frailty models have been introduced into the statistical literature in an attempt to account for the existence of heterogeneity in a population

*Corresponding author: e-mail: victorleivasanchez@gmail.com; Phone: +56 32 2503815; URL: www.victorleiva.cl

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

under study. The random variable (RV) "frailty" may be incorporated in the baseline hazard rate (HR) additively or multiplicatively. Several authors have considered frailty models, which represent a generalization of the Cox model; see Cox (1972) and Stare and O'Quigley (2004). The interested reader in frailty models is referred to Hougaard (2000), Duchateau and Janssen (2008), and Wienke (2011). These works present reviews of frailty models and inferential procedures from both classical and Bayesian perspectives. Some other works about frailty models are the following. Aalen and Tretli (1999) studied incidence of testicular cancer with frailty models. Henderson and Oman (1999) detected the consequence of ignoring the frailty on marginal regression estimates in survival models. Fan and Li (2002) considered variable selection in frailty models. Cai (2010) analyzed Bayesian semiparametric frailty selection in multivariate event time data. Mazroui et al. (2013) proposed a multivariate frailty model that jointly assumes two types of recurrent events with a dependent terminal event.

Because of the randomness of the frailty term, it is necessary to consider a distribution for it. Moreover, due to the way how the frailty term acts on the HR, natural candidates to the frailty distribution are the gamma (GA), inverse Gaussian (IG), lognormal (LN), and Weibull models. Especially, from the seminal work presented by Vaupel et al. (1979), the GA frailty distribution has been used in most applications published up to date; see Balakrishnan and Peng (2006).

A good alternative to the GA distribution is the Birnbaum–Saunders (BS) distribution. It has been widely considered in the literature due to its physical arguments, its attractive properties, and its relationship with the normal distribution. The BS model was proposed by Birnbaum and Saunders (1969) and has been extensively applied for modeling failure times in engineering, although some novel applications have been considered in biological, environmental, and financial sciences; see, for example, Desmond (1985), Kotz et al. (2010), Saulo et al. (2013), and Leiva et al. (2014b, 2014c, 2015a, 2015b, 2017). Santos-Neto et al. (2012) introduced several parameterizations of the BS distribution. Specially, one of them is established in terms of the distribution mean, whereas its variance is a quadratic function of this mean. Thus, such a parameterization allows us to mimic a property of the GA frailty distribution early proposed by Vaupel et al. (1979), doing the BS frailty distribution to be a new alternative to frailty modeling. In this paper, we propose a new frailty regression model based on the BS distribution.

Diagnostic analysis is an efficient way to detect influential cases and evaluate their effect on the model inference. To the best of our knowledge, influence diagnostic tools have not been considered in frailty models. The natural way to assess the effect of an observation on the estimation is case deletion, which is considered as a global influence technique. However, it excludes all data from a case and can be a problem to know whether that case has some influence on a specific aspect of the model or not. To overcome this problem, one may use the local influence technique; see Cook (1987). It allows us to detect locally influential cases and provides a sensitivity measure under perturbations on the data or the model. The local influence technique has been extended to various regression models; see, for example, Osorio et al. (2007), Espinheira et al. (2008), Paula et al. (2009), and Leiva et al. (2014a). Note that global and local influence techniques can be based on the generalized Cook distance (GCD) and the likelihood function, respectively.

The main objectives of this paper are: (i) to propose a BS frailty regression model and its inference based on the maximum-likelihood (ML) method, and (ii) to derive influence diagnostic tools for this model. The secondary objective is to apply the BS frailty regression model and its diagnostics to medical data to illustrate its potential applications and compare it with classical frailty models.

Section 2 defines the frailty model, discusses how to obtain the HR and survival function (SF), and presents properties of the BS distribution. The model parameters estimated via the ML method and inference are both provided in Section 3. Global influence and the normal curvatures of local influence under different perturbation schemes are derived in Section 4. Also, two types of residuals for the new BS frailty regression model are introduced. In Section 5, we illustrate the proposed model and its diagnostics with two medical real-world data, comparing it to classical frailty models. Discussion, some concluding remarks, and possible future studies are detailed in Section 6.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

2 Background

In this section, we provide some preliminary aspects of frailty models, HR and SF, and present some features of the BS distribution.

2.1 Frailty models

Consider an unobserved source of heterogeneity that is not readily captured by a covariate in a univariate frailty model. This extends the Cox model, such that the HR of a patient depends on an unobservable RV U, which acts multiplicatively on the baseline HR. Therefore, the conditional HR of the lifetime T, given $U = u_i$ for the patient i at time t, is

$$h_{T|U=u_i}(t; \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = u_i h_0(t; \boldsymbol{\xi}_1), \quad i = 1, \ldots, n, \quad t > 0, \tag{1}$$

where u_i is the frailty of the patient i and h_0 is a baseline HR, that is, we consider the case with a proportional HR. In (1), note that $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are vectors of the model parameters related to the lifetime and frailty distributions, respectively, of the patient i. In addition, observe that (1) is known as the Clayton model; see Clayton (1991). From (1), a patient i is called "standard" if his/her frailty is $u_i = 1$; "twice as likely to die" if his/her frailty is $u_i = 2$, at any particular time and in relation to the standard patient; and "one-half as likely to die" if his/her frailty is $u_i = 1/2$; see Vaupel et al. (1979). The corresponding conditional SF of T is $S_{T|U=u_i}(t; \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = (S_0(t; \boldsymbol{\xi}_1))^{u_i}$, for $i = 1, \ldots, n$ and $t > 0$, which represents the probability of the patient i to be alive at time t given the random effect $U_i = u_i$.

If values of covariates in the model given in (1) are introduced similarly to the Cox model, we have

$$h_{T|U=u_i}(t; \boldsymbol{x}, \boldsymbol{\xi}) = u_i h_0(t; \boldsymbol{\xi}_1) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \quad i = 1, \ldots, n, \quad t > 0, \tag{2}$$

where $\boldsymbol{x}_i^\top = (1, x_{1i}, \ldots, x_{pi})$ is a vector containing the values of p covariates for the patient i, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the vector of regression coefficients to be estimated, and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\top, \boldsymbol{\xi}_2^\top, \boldsymbol{\beta}^\top)^\top$. Therefore, the frailty model given in (2) is a generalization of the proportional hazard model, which is obtained when the frailty distribution degenerates at $U = 1$ for all patients. The corresponding conditional SF can be obtained from (2) as

$$S_{T|U=u_i}(t; \boldsymbol{x}, \boldsymbol{\xi}) = \exp(-u_i H_0(t; \boldsymbol{\xi}_1) \exp(\boldsymbol{x}^\top \boldsymbol{\beta})), \quad i = 1, \ldots, n, \quad t > 0, \tag{3}$$

where $H_0(t; \boldsymbol{\xi}_1) = \int_0^t h_0(s; \boldsymbol{\xi}_1)\, \mathrm{d}s$ is the baseline cumulative HR (CHR).

Suppose that the lifetime is not completely observed and may be subject to right censoring. Let v_i denote the censoring time, y_i the time to event of interest, and u_i the frailty for the patient i, respectively. We observe $t_i = \min\{y_i, v_i\}$, that is, if the censoring indicator $\tau_i = 1$, $t_i = y_i$ is the lifetime of the patient i; otherwise, if $\tau_i = 0$, $t_i = v_i$ is the right censoring time of the patient i; for $i = 1, \ldots, n$. Then, from (2) and (3), the corresponding likelihood function for $\boldsymbol{\xi}$ is

$$L(\boldsymbol{\xi}; \boldsymbol{t}, \boldsymbol{\tau}, \boldsymbol{x}, \boldsymbol{u}) = L(\boldsymbol{\xi}) = \prod_{i=1}^{n} \left(u_i h_0(t; \boldsymbol{\xi}_1) \exp(\boldsymbol{x}^\top \boldsymbol{\beta})\right)^{\tau_i} \exp\left(-u_i H_0(t; \boldsymbol{\xi}_1) \exp(\boldsymbol{x}^\top \boldsymbol{\beta})\right), \tag{4}$$

where $\boldsymbol{\xi}$ is defined in (2), $\boldsymbol{t} = (t_1, \ldots, t_n)^\top$ are the lifetimes of the patients, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^\top$ is the vector of their censoring indicators, and $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$ is the vector of their frailties. Now, conditional on the unobserved frailties $\boldsymbol{u}$, the likelihood function given in (4) forms the basis for the parameter estimation. The frailties $\boldsymbol{u}$ must be integrated out (in closed form or by numerical or stochastic

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

integration, depending on the frailty distribution) to get a likelihood function for $\boldsymbol{\xi}$ (not depending on unobserved quantities) of the type

$$L(\boldsymbol{\xi}; \boldsymbol{t}, \boldsymbol{\tau}, \boldsymbol{x}) = L(\boldsymbol{\xi}) = \prod_{i=1}^{n} (h_T(t_i; \boldsymbol{x}, \boldsymbol{\xi}))^{\tau_i} S_T(t_i; \boldsymbol{x}, \boldsymbol{\xi}), \tag{5}$$

where h_T and S_T are the unconditional HR and SF, respectively, defined next.

2.2 Unconditional HR and SF

The unconditional (population) SF of T can be obtained by integrating $S_{T|U=u_i}(t; \boldsymbol{x})$ given in (3) on the frailty U. It may be viewed as the (unconditional) SF of patients randomly drawn from the population under study; see Klein and Moeschberger (2003), Aalen et al. (2008), and Wienke (2011). Unconditional HF and SF may be get with the Laplace transform; see Hougaard (1984). Then, when seeking distributions for the frailty RV U, it is natural to use frailty distributions with an explicit Laplace transform, because it facilitates the use of standard ML methods for parameter estimation. To get the unconditional SF, we need to integrate out the frailty component as

$$S_T(t; \boldsymbol{x}, \boldsymbol{\xi}) = \int_0^\infty S_{T|U=u}(t; \boldsymbol{x}, \boldsymbol{\xi}) f_U(u; \boldsymbol{\xi}_2)\, \mathrm{d}u, \tag{6}$$

where $\boldsymbol{\xi}$ is defined in (2), $S_{T|U=u}(t; \boldsymbol{x}, \boldsymbol{\xi}) = \exp(-uH_0(t; \boldsymbol{\xi}_1)\exp(\boldsymbol{x}^\top \boldsymbol{\beta}))$ is the conditional SF as given in (3), and f_U is the corresponding frailty probability density function (PDF). The Laplace transform of real argument s of a function f is

$$Q(s) = \int_0^\infty \exp(-sx) f(x)\, \mathrm{d}x. \tag{7}$$

Let $f = f_U$ be the frailty PDF and $s = H_0(t; \boldsymbol{\xi}_1)\exp(\boldsymbol{x}^\top \boldsymbol{\beta})$. Then, according to (7), we obtain the Laplace transform of the unconditional SF of T as

$$S_T(t; \boldsymbol{x}, \boldsymbol{\xi}) = \int_0^\infty \exp(-uH_0(t; \boldsymbol{\xi}_1)\exp(\boldsymbol{x}^\top \boldsymbol{\beta})) f_U(u; \boldsymbol{\xi}_2)\, \mathrm{d}u = Q(H_0(t; \boldsymbol{\xi}_1)\exp(\boldsymbol{x}^\top \boldsymbol{\beta})). \tag{8}$$

Note that (8) conducts to the same form as the unconditional SF given in (6); see Vaupel et al. (1979) and Wienke (2011). The frailty RVs U_i are usually assumed to be independent with identical frailty distribution. As mentioned, the frailty distribution can be GA, IG, LN, or Weibull. We consider a reparameterized version of the BS distribution introduced by Santos-Neto et al. (2012, 2014), because it allows us to mimic a property of the GA distribution, traditionally used in frailty models.

2.3 A BS distribution

Santos-Neto et al. (2012) proposed several parameterizations of the BS distribution, which allow diverse features of data modeling to be considered. One of such parameterizations is indexed by the parameters $\mu = \beta(1 + \alpha^2/2)$ and $\delta = 2/\alpha^2$, where $\alpha > 0$ and $\beta > 0$ are the original BS parameters (see Birnbaum and Saunders, 1969), $\mu > 0$ is a scale parameter and the mean of the distribution, whereas $\delta > 0$ is a shape and precision parameter. The notation $U \sim \mathrm{BS}(\mu, \delta)$ is used when the RV U follows such a distribution. This parameterization of the BS distribution permits us to mimic a property of the GA distribution, which was the first distribution used in a frailty model (see Vaupel et al., 1979), as follows. The mean and variance of $U \sim \mathrm{BS}(\mu, \delta)$ are $\mathrm{E}[U] = \mu$ and $\mathrm{Var}[U] = \mu^2/\phi$, respectively, where $\phi = (\delta + 1)^2/(2\delta + 5)$. Then, as mentioned, δ can be interpreted as a precision parameter, that is, for fixed values of μ, when $\delta \to \infty$, the variance of T tends to zero. Also, for fixed μ, if $\delta \to 0$, then $\mathrm{Var}[U] \to 5\mu^2$. Note that $\mathrm{Var}[U] = \mu^2/\phi$ is similar to the variance function of the GA distribution,

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

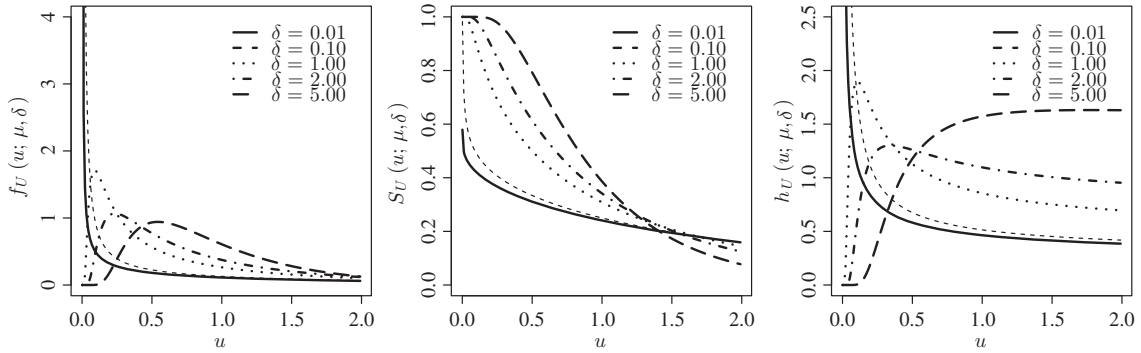

Figure 1 PDF (left), SF (center), and HR (right) plots of the BS model for $\mu = 1$ and indicated values of δ.

which has a quadratic relation with its mean. Therefore, a frailty model based on the BS distribution, in its reparameterized form, can be a good alternative to the GA frailty model.

If $U \sim \text{BS}(\mu, \delta)$, then its PDF is

$$f_U(u;\mu,\delta) = \frac{\exp(\delta/2)\sqrt{\delta+1}}{4\,u^{\frac{3}{2}}\sqrt{\pi\mu}}\left(u+\frac{\delta\mu}{\delta+1}\right)\exp\left(-\frac{\delta}{4}\left(\frac{u(\delta+1)}{\delta\mu}+\frac{\delta\mu}{u(\delta+1)}\right)\right), \quad u>0. \tag{9}$$

It is possible to show that $k\,U \sim \text{BS}(k\,\mu, \delta)$, with $k > 0$, and $1/U \sim \text{BS}(\mu^*, \delta)$, where $\mu^* = (\delta + 1)/(\delta\mu)$, that is, the BS distribution, in its original and reparameterized forms, is closed under scaling and reciprocation. From (9), the SF and HR of U are, respectively,

$$S_U(u;\mu,\delta) = \frac{1}{2}\Phi\left(u+\delta(u-\mu)/(2\sqrt{u(1+\delta)\mu})\right), \quad u>0,$$

$$h_U(u;\mu,\delta) = \frac{\exp(-(-\delta\mu+\delta u+u)^2/(4(\delta+1)\mu u))(\delta\mu+\delta u+u)}{(\pi\mu(\delta+1))^{\frac{1}{2}}2\mu^{\frac{1}{2}}u^{\frac{3}{2}}\Phi((u+\delta(u-\mu))/(2\sqrt{u(1+\delta)\mu}))}, \quad u>0,$$

where Φ is the N(0, 1) cumulative distribution function (CDF). Figure 1 displays some shapes for the PDF, SF, and HR of $U \sim \text{BS}(\mu = 1, \delta)$. Note that a unimodal behavior is detected for the PDF, as well as different degrees of asymmetry and kurtosis, whereas the HR has increasing and decreasing shapes, such as the GA distribution, but also an inverse bathtub shape.

The use of the BS distribution has the following appealing advantages:

- Based on its genesis, it is possible to make an analogy in the modeling of medical data; see, for example, Desmond (1985).
- Its parameterization based on the mean (μ), such as in (9), allows us to analyze data in their original scale, avoiding, for instance, problems of interpretation in models that employ a logarithmic transformation of the data; see Leiva et al. (2014b) and Santos-Neto et al. (2014, 2016).
- In the context of frailty models, it can be very competitive in terms of fitting.
- It belongs to the class of log-symmetric distributions, such as the case of the generalized BS, LN, log-logistic, log-Laplace, log-Student-t, log-power-exponential, log-slash, and F distributions; see Vanegas and Paula (2016a, 2016b). The log-symmetric class of distributions arises when an RV has the same distribution as its reciprocal or as ordinary symmetry of the distribution of the logged RV; see Jones (2008). One can obtain the BS, LN, and log-logistic frailty models as particular

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

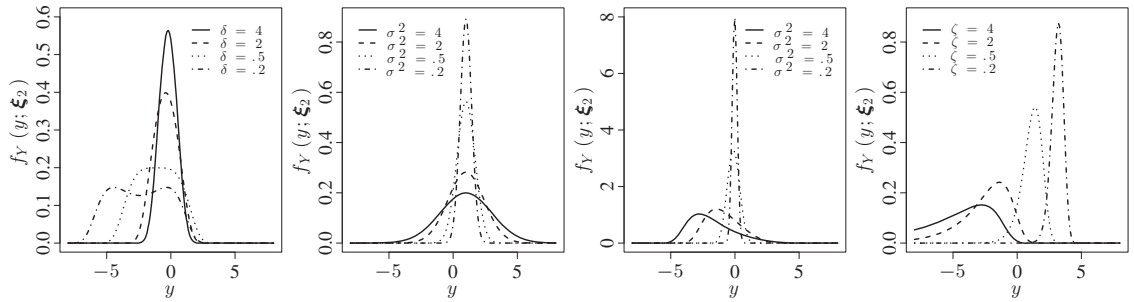

Figure 2 PDF plots of the log-BS($\sqrt{2/\delta}$, $\log(\delta/(\delta+1))$) (left), N($1, \sigma^2$) (left-center), log-IG($\sigma^2, 0$) (right-center), and log-GA($1/\zeta, 1/\zeta$) (right) distributions.

cases of log-symmetric frailty models. However, besides the BS frailty model that is proposed in this paper, the only other popular log-symmetric frailty model that belongs to this class is the LN one. But unlike the BS frailty model, the LN model does not have an explicit Laplace transform; see Wienke (2011). This explicit form is useful to obtain the PDF and the unconditional SF and HR.

- It is flexible in terms of bimodality when the logarithm of a BS RV is taken into account. Note that

  (i) If $U \sim \text{BS}(1, \delta)$, $Y = \log(U) \sim \text{log-BS}(\sqrt{2/\delta}, \log(\delta/(\delta+1)))$; see Rieck and Nedelman (1991);
  (ii) If $U \sim \text{LN}(1, \sigma^2)$, $Y = \log(U) \sim \text{N}(1, \sigma^2)$; see Crow and Shimizu (1988);
  (iii) If $U \sim \text{IG}(1, \sigma^2)$, $Y = \log(U) \sim \text{log-IG}(\sigma^2, 0)$; see Kotz et al. (2010);
  (iv) If $U \sim \text{GA}(1/\zeta, 1/\zeta)$, $Y = \log(U) \sim \text{log-GA}(1/\zeta, 1/\zeta)$; see Johnson et al. (1995).

  Some properties of the log-BS distribution are as follows. If $Y \sim \text{log-BS}(\sqrt{2/\delta}, \log(\delta\mu/(\delta+1)))$, then (a) $U = \exp(Y) \sim \text{BS}(\mu, \delta)$; (b) $\text{E}(Y) = \log(\delta\mu/(\delta+1))$; (c) there is no closed form for the variance of Y, but based upon an asymptotic approximation for the log-BS moment generating function, it follows that, as $\delta \to \infty$, $\text{Var}(Y) = 2/\delta - 1/\delta^2$, whereas that, in contrast, as $\delta \to 0$, $\text{Var}(Y) = 4(\log^2(2/\sqrt{\delta}) + 2 - 2\log(2/\sqrt{\delta}))$; and (d) the distribution of Y is symmetric around μ, unimodal for $\delta \geq 0.5$, and bimodal for $\delta < 0.5$; see Rieck and Nedelman (1991) and Leiva (2016). Figure 2 shows some shapes for the PDF of $Y = \log(U)$ in each aforementioned distribution. Note that the bimodality property is only found in the log-BS case.

3 BS frailty regression model

In this section, we discuss some model identifiability issues and how to estimate the model parameters via the ML method and to infer about these parameters.

3.1 Model identifiability and features

In frailty models, an important aspect to be studied is its identifiability. In the context of proportional hazard models, when working with frailty, it is necessary that the random effect distribution has finite mean for the model to be identifiable; see Elbers and Ridder (1982). Thus, in order to keep the identifiability of the frailty model, it is convenient to have a distribution with mean equal to 1. We assume that the frailty U has a BS distribution with parameters $\mu = 1$ and δ, where $\text{E}[U] = 1$ and $\text{Var}[U] = (2\delta + 5)/(\delta + 1)^2$. The variance quantifies the amount of heterogeneity among patients.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

From (7), the Laplace transform for the BS distribution with parameters $\mu = 1$ and δ is

$$Q(s) = \frac{\exp\left(\frac{\delta}{2}\left(1 - \sqrt{\delta + 4s + 1}/\sqrt{\delta + 1}\right)\right)\left(\sqrt{\delta + 4s + 1} + \sqrt{\delta + 1}\right)}{2\sqrt{\delta + 4s + 1}}. \tag{10}$$

From (8) and evaluating (10) at $s = H_0(t; \boldsymbol{\xi}_1)\exp(\eta)$, with $\eta = \boldsymbol{x}^\top \boldsymbol{\beta}$, we obtain the unconditional SF of T under the BS frailty as

$$S_T(t; \boldsymbol{x}, \boldsymbol{\xi}) = \frac{\exp\left(\frac{\delta}{2}\left(1 - \frac{\sqrt{\delta + 4H_0(t;\boldsymbol{\xi}_1)\exp(\eta)+1}}{\sqrt{\delta+1}}\right)\right)\left(\sqrt{\delta + 4H_0(t; \boldsymbol{\xi}_1)\exp(\eta) + 1} + \sqrt{\delta + 1}\right)}{2\sqrt{\delta + 4H_0(t; \boldsymbol{\xi}_1)\exp(\eta) + 1}}. \tag{11}$$

Then, from (10) and (11), the corresponding unconditional HR of T is

$$\begin{aligned} h_T(t; \boldsymbol{x}, \boldsymbol{\xi}) &= h_0(t; \boldsymbol{\xi}_1)\exp(\eta) \times \\ &\times \frac{\delta(\delta + \sqrt{\delta + 1}\sqrt{\delta + 4H_0(t; \boldsymbol{\xi}_1)\exp(\eta) + 1} + 4H_0(t; \boldsymbol{\xi}_1)\exp(\eta) + 3) + 2}{(\delta + 4H_0(t; \boldsymbol{\xi}_1)\exp(\eta) + 1)(\delta + \sqrt{\delta + 1}\sqrt{\delta + 4H_0(t; \boldsymbol{\xi}_1)\exp(\eta) + 1} + 1)}. \end{aligned} \tag{12}$$

We assume that h_0 is specified up to a few unknown parameters, which are related to a distribution assumed for the baseline HR. For example, we can suppose an exponential, LN, or Weibull distribution, among others. However, assuming a parametric distribution is not always desirable, because such an assumption is often difficult to verify. Note that the Weibull distribution has been extensively used to model the baseline HR due to its flexibility or when the HR must be constant, increasing, or decreasing for each patient; see Lawless (2003). Therefore, we use the Weibull distribution as baseline hazard, which has $h_0(t; \gamma, \kappa) = \gamma\kappa t^{\kappa-1}$ and $H_0(t; \gamma, \kappa) = \gamma t^\kappa$, for $t > 0$, where $\kappa > 0$ and $\gamma > 0$ are shape and scale parameters, respectively. Note that the baseline HR h_0: (i) increases if $\kappa > 1$; (ii) is constant (exponential model) if $\kappa = 1$; and (iii) decreases if $\kappa < 1$. This parameterization is commonly used in statistical models for medicine; see Collett (2015). Thus, from (12) and for $\boldsymbol{\xi} = (\gamma, \kappa, \delta, \eta)^\top$, with $\eta = \boldsymbol{x}^\top\boldsymbol{\beta}$ as given in (11), the unconditional HR and SF of T under BS frailty reduce to

$$\begin{aligned} h_T(t; \boldsymbol{x}, \boldsymbol{\xi}) &= \frac{\gamma\kappa t^{\kappa-1}\exp(\eta)\,(\delta(\delta + \Delta(t; \boldsymbol{\xi}) + 4\gamma t^\kappa\exp(\eta) + 3) + 2)}{\Delta^2(t; \boldsymbol{\xi})(\delta + \Delta(t; \boldsymbol{\xi}) + 1)}, \\ S_T(t; \boldsymbol{x}, \boldsymbol{\xi}) &= \frac{\exp((\delta/2)(1 - \Delta^*(t; \boldsymbol{\xi})/\sqrt{\delta + 1}))(\Delta^*(t; \boldsymbol{\xi}) + \sqrt{\delta + 1})}{2\Delta^*(t; \boldsymbol{\xi})}, \end{aligned} \tag{13}$$

where $\Delta(t; \boldsymbol{\xi}) = \sqrt{(\delta + 1)(\delta + 4\gamma t^\kappa\exp(\eta) + 1)}$ and $\Delta^*(t; \boldsymbol{\xi}) = \sqrt{\delta + 4\gamma t^\kappa\exp(\eta) + 1}$.

3.2 Estimation of parameters

Let n patients provide pairs of lifetimes and right censoring indicators (t_i, τ_i), for $i = 1, \ldots, n$, with t_i and τ_i being the elements of the vectors $\boldsymbol{t}$ and $\boldsymbol{\tau}$ defined in (4), respectively. In addition, consider the BS frailty regression model given in (13) with parameter vector $\boldsymbol{\xi} = (\gamma, \kappa, \delta, \boldsymbol{\beta}^\top)^\top$. Then, the

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

corresponding log-likelihood function for $\boldsymbol{\xi}$ under uninformative censoring, taken from (5), can be expressed as

$$\ell(\boldsymbol{\xi}) = \log(L(\boldsymbol{\xi})) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\xi}), \tag{14}$$

where

$$\begin{aligned}\ell_i(\boldsymbol{\xi}) = {} & \tau_i(\log(\kappa) + \log(\gamma) + (\kappa - 1)\log(t_i) + \eta) + (\delta/2)(1 - (\Delta^*(t_i; \boldsymbol{\xi}))/\sqrt{\delta + 1}) + \\ & + \log(\Delta^*(t_i; \boldsymbol{\xi}) + \sqrt{\delta + 1}) - \log(2\Delta^*(t_i; \boldsymbol{\xi})) + \\ & + \tau_i \log(\delta(\delta + \Delta(t_i; \boldsymbol{\xi}) + 4\gamma t_i^{\kappa} \exp(\eta) + 3) + 2) - 2\tau_i \log(\Delta^*(t_i; \boldsymbol{\xi})(\delta + \Delta(t_i; \boldsymbol{\xi}) + 1)),\end{aligned}$$

with $\eta = \boldsymbol{x}^\top \boldsymbol{\beta}$ defined in (11). The parameter vector $\boldsymbol{\xi}$ may be estimated by numerical maximization of the log-likelihood function ℓ given in (14) using the `R` software by its functions `optim` and `optimx`; see www.R-project.org and R-Team (2016).

3.3 Inference and simulation

It can be verified that standard regularity conditions (see Cox and Hinkley, 1974) are fulfilled for the proposed model, whenever the parameters are within the parameter space. Then, the ML estimator $\widehat{\boldsymbol{\xi}}$ is consistent and follows a normal asymptotic joint distribution with asymptotic mean $\boldsymbol{\xi}$, and an asymptotic covariance matrix $\Sigma(\widehat{\boldsymbol{\xi}})$ that can be obtained from the corresponding expected Fisher information matrix. Thus, recalling that $\boldsymbol{\xi} = (\gamma, \kappa, \delta, \boldsymbol{\beta}^\top)^\top$, we have, as $n \to \infty$,

$$\sqrt{n}\,(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}) \quad \stackrel{D}{\to} \quad \mathrm{N}_{p+4}(\mathbf{0}_{(p+4)\times 1}, \Sigma(\widehat{\boldsymbol{\xi}}) = \mathcal{J}(\boldsymbol{\xi})^{-1}), \tag{15}$$

where $\mathbf{0}_{(p+4)\times 1}$ is a $(p+4) \times 1$ vector of zeros, $\mathcal{J}(\boldsymbol{\xi}) = \lim_{n\to\infty}(1/n)\mathcal{I}(\boldsymbol{\xi})$, with $\mathcal{I}(\boldsymbol{\xi})$ being the corresponding expected Fisher information matrix, and $\stackrel{D}{\to}$ denotes convergence in distribution. Note that $\widehat{\mathcal{I}}(\boldsymbol{\xi})^{-1}$ is a consistent estimator of the variance–covariance matrix of $\widehat{\boldsymbol{\xi}}$, $\Sigma(\widehat{\boldsymbol{\xi}})$ namely. In practice, one may approximate the expected Fisher information matrix by its observed version, whereas the diagonal elements of its inverse matrix can be used to approximate the corresponding standard errors (SEs); see Efron and Hinkley (1978) for the use of observed versus expected Fisher information matrices. Besides estimation, hypothesis testing is another important topic to be addressed. Let $\boldsymbol{\xi}^*$ be a proper disjoint subset of $\boldsymbol{\xi}$. We aim to test H_0: $\boldsymbol{\xi}^* = \boldsymbol{\xi}_0^*$ versus H_1: $\boldsymbol{\xi}^* \neq \boldsymbol{\xi}_0^*$. Also, let $\widehat{\boldsymbol{\xi}}_0^*$ maximize $\ell(\boldsymbol{\xi}^*)$ given in (14) constrained to H_0. Then, the corresponding likelihood ratio (LR) statistic is $\mathrm{LR} = 2\log(\ell(\widehat{\boldsymbol{\xi}}^*)/\ell(\widehat{\boldsymbol{\xi}}_0^*))$. Under H_0 and some regularity conditions, that is, conditions needed for the asymptotic theory of ML estimators to hold (see Serfling, 1980), the distribution of the LR statistic converges to the $\chi^2(\varrho)$ distribution, with $\varrho = \dim(\boldsymbol{\xi}^*)$ being the dimension of the vector $\boldsymbol{\xi}^*$.

Simulation studies to evaluate the performance of the ML estimators can be carried out by generating random numbers from the BS frailty regression model from Algorithm 1. Without loss of generality, one covariate can be assumed.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Algorithm 1. Generator of random numbers from the BS frailty regression model.

1: Obtain a random number x_i from $X \sim \mathrm{U}(0, 1)$.
2: Set a value for $\boldsymbol{\xi} = (\gamma, \kappa, \delta, \beta)^\top$.
3: Calculate $\eta_i = x_i\beta$ and determine expressions for $\Delta^*(y_i; \boldsymbol{\xi})$ and $\Delta(y_i; \boldsymbol{\xi})$ as a function of y_i based on the formulas given in (13).
4: Generate a random number m_i from $M \sim \mathrm{U}(0, 1)$.
5: Equate m_i to the SF defined in (13) and get the time to event of interest y_i by solving numerically the equation

$$m_i = \frac{\exp\left(\frac{\delta}{2}\left(1 - \Delta^*(y_i; \boldsymbol{\xi})/\sqrt{\delta+1}\right)\left(\Delta^*(y_i; \boldsymbol{\xi}) + \sqrt{\delta+1}\right)\right.}{2\Delta^*(y_i; \boldsymbol{\xi})}.$$

6: Establish the censoring time v_i from $V \sim \mathrm{U}(a, b)$, for $a > 0$ and $b > 0$ fixed.
7: Find a random number $t_i = \min\{y_i, v_i\}$, that is,
7.1: If $y_i < v_i$, then $\tau_i = 1$ and $t_i = y_i$;
7.2: Else, $\tau_i = 0$ and $t_i = v_i$.
8: Repeat Steps 1 to 7 until the amount of n random numbers to be completed.

4 Influence diagnostics and residual analysis

In this section, we introduce global of local influence techniques and two types of residuals for the BS frailty regression model.

4.1 Global influence

Global influence is related to case deletion, that is, it is a technique to study the effect of dropping a case from the data set. Consider a version for case deletion from the expressions given in (13), with the subscript "(i)" meaning the set of patients has the case i deleted. Consequently, the corresponding log-likelihood function defined in (14) is now denoted by $\ell_{(i)}$. Let $\widehat{\boldsymbol{\xi}}_{(i)} = (\widehat{\gamma}_{(i)}, \widehat{\kappa}_{(i)}, \widehat{\delta}_{(i)}, \widehat{\boldsymbol{\beta}}_{(i)}^\top)^\top$ be the ML estimate of $\boldsymbol{\xi}$ from $\ell_{(i)}$. To assess the influence of the case i on the ML estimate $\widehat{\boldsymbol{\xi}} = (\widehat{\gamma}, \widehat{\kappa}, \widehat{\delta}, \widehat{\boldsymbol{\beta}}^\top)^\top$, the basic idea is to compare the difference between $\widehat{\boldsymbol{\xi}}_{(i)}$ and $\widehat{\boldsymbol{\xi}}$ in terms of $\ell_{(i)}$ and ℓ, respectively. If deletion of a case seriously influences the estimates, more attention should be paid to that case. Hence, if $\widehat{\boldsymbol{\xi}}_{(i)}$ is far from $\widehat{\boldsymbol{\xi}}$, the case i is regarded as potentially influential. A first measure of global influence is defined as the standardized norm of $\widehat{\boldsymbol{\xi}}_{(i)} - \widehat{\boldsymbol{\xi}}$, known as the GCD, given by $\mathrm{GCD}_i(\boldsymbol{\xi}) = (\widehat{\boldsymbol{\xi}}_{(i)} - \widehat{\boldsymbol{\xi}})^\top(\widehat{\Sigma}(\widehat{\boldsymbol{\xi}}))^{-1}(\widehat{\boldsymbol{\xi}}_{(i)} - \widehat{\boldsymbol{\xi}})$, where $\widehat{\Sigma}(\widehat{\boldsymbol{\xi}})$ is an estimate of $\Sigma(\widehat{\boldsymbol{\xi}})$ obtained from (15). An alternative way is to assess $\mathrm{GCD}_i(\gamma)$, $\mathrm{GCD}_i(\kappa)$, $\mathrm{GCD}_i(\delta)$, and $\mathrm{GCD}_i(\boldsymbol{\beta})$, whose values reveal the impact of the case i on the estimates of γ, κ, δ, and $\boldsymbol{\beta}$, respectively. Also, $\widehat{\boldsymbol{\xi}}_{(i)}$ and $\widehat{\boldsymbol{\xi}}$ can be compared by their likelihood distance (LD) defined as $\mathrm{LD}_i(\boldsymbol{\xi}) = 2(\ell(\widehat{\boldsymbol{\xi}}) - \ell(\widehat{\boldsymbol{\xi}}_{(i)}))$, for $i = 1, \ldots, n$.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

4.2 Local influence

Local influence is based on the curvature of the plane of the log-likelihood function. Consider the BS frailty regression model given in (13), recall $\boldsymbol{\xi} = (\gamma, \kappa, \delta, \boldsymbol{\beta}^\top)^\top$ and let $\ell(\boldsymbol{\xi}; \boldsymbol{\omega})$ be the log-likelihood function corresponding to this model defined in (14) but now perturbed by $\boldsymbol{\omega}$. The vector of perturbations $\boldsymbol{\omega}$ belongs to a subset $\Omega \in \mathbb{R}^n$ and $\boldsymbol{\omega}_0$ is a nonperturbed $n \times 1$ vector, such that $\ell(\boldsymbol{\xi}; \boldsymbol{\omega}_0) = \ell(\boldsymbol{\xi})$, for all $\boldsymbol{\xi}$. In this case, the LD is $\mathrm{LD}(\boldsymbol{\xi}) = 2(\ell(\widehat{\boldsymbol{\xi}}) - \ell(\widehat{\boldsymbol{\xi}}_\omega))$, where $\widehat{\boldsymbol{\xi}}_\omega$ denotes the ML estimate of $\boldsymbol{\xi}$ upon the perturbed BS frailty regression model, that is, $\widehat{\boldsymbol{\xi}}_\omega$ is obtained from $\ell(\boldsymbol{\xi}; \boldsymbol{\omega})$. Note that $\ell(\boldsymbol{\xi}; \boldsymbol{\omega})$ can be used to assess the influence of the perturbation on the ML estimate. Cook (1987) showed that the normal curvature for $\widehat{\boldsymbol{\xi}}$ in the direction $\boldsymbol{d}$, with $||\boldsymbol{d}|| = 1$, is expressed as $C_{\boldsymbol{d}}(\widehat{\boldsymbol{\xi}}) = 2|\boldsymbol{d}^\top \nabla^\top \Sigma(\widehat{\boldsymbol{\xi}})^{-1} \nabla \boldsymbol{d}|$, where ∇ is a $(p+4) \times n$ matrix of perturbations with elements $\nabla_{ji} = \partial^2 \ell(\xi; \boldsymbol{\omega}) / \partial \xi_j \partial \omega_i$, evaluated at $\boldsymbol{\xi} = \widehat{\boldsymbol{\xi}}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, for $j = 1, \ldots, p+4$ and $i = 1, \ldots, n$. A local influence diagnostic is generally based on index plots. For example, the index graph of the eigenvector $\boldsymbol{d}_{\max}$ related to the maximum eigenvalue of $B(\boldsymbol{\xi}) = -\nabla^\top \Sigma(\boldsymbol{\xi})^{-1} \nabla$, say $C_{d_{\max}}(\boldsymbol{\xi})$, evaluated at $\boldsymbol{\xi} = \widehat{\boldsymbol{\xi}}$, can detect those cases that, under small perturbations, exercise a high influence on $\mathrm{LD}(\boldsymbol{\xi})$.

Another important direction of interest is $\boldsymbol{d}_i = \boldsymbol{e}_{in}$, which corresponds to the direction of the case i, where $\boldsymbol{e}_{in}$ is an $n \times 1$ vector of zeros with a value equal to 1 at the i-th position, that is, $\{\boldsymbol{e}_{in}, 1 \le i \le n\}$ is the canonical basis of $\mathbb{R}^n$. In this case, the normal curvature is $C_i(\boldsymbol{\xi}) = 2|b_{ii}|$, where b_{ii} is the i-th diagonal element of $B(\boldsymbol{\xi})$ given above, for $i = 1, \ldots, n$, evaluated at $\boldsymbol{\xi} = \widehat{\boldsymbol{\xi}}$. If $C_i(\widehat{\boldsymbol{\xi}}) > 2\overline{C}(\widehat{\boldsymbol{\xi}})$, where $\overline{C}(\widehat{\boldsymbol{\xi}}) = \sum_{i=1}^n C_i(\widehat{\boldsymbol{\xi}})/n$, it indicates the case i as potentially influential. This procedure is called total local influence of the case i and can be carried for $\boldsymbol{\xi}$ or for γ, κ, δ, or $\boldsymbol{\beta}$, which are denoted as $C_i(\boldsymbol{\xi})$, $C_i(\gamma)$, $C_i(\kappa)$, $C_i(\delta)$, and $C_i(\boldsymbol{\beta})$, respectively.

We consider the model defined in (13) and its log-likelihood function given by (14). Due to space restrictions, we omit the elements $\nabla(\gamma)$, $\nabla(\kappa)$, $\nabla(\delta)$, and $\nabla(\boldsymbol{\beta}^\top)$ of the matrix ∇ for each perturbation scheme detailed below.

4.2.1 Case-weight perturbation

Under this perturbation scheme, we evaluate whether the contributions of the cases with different weights affect the ML estimate of $\boldsymbol{\xi}$. The log-likelihood function of the perturbed BS frailty model is $\ell(\boldsymbol{\xi}; \boldsymbol{\omega}) = \sum_{i=1}^n \omega_i \ell_i(\boldsymbol{\xi})$, where $0 \le \omega_i \le 1$, $\boldsymbol{\omega}_0 = (1, \ldots, 1)^\top$, and $\ell_i(\boldsymbol{\xi})$ given in (14).

4.2.2 Response perturbation

We here assume an additive perturbation on the response variable (lifetime) for the case i such that $t_i(\omega_i) = t_i + \omega_i s_T$, where $s_T = (1/\widehat{\phi})^{1/2}$ is a scale factor and $\omega_i \in \mathbb{R}$, for $i = 1, \ldots, n$. Then, the log-likelihood function is $\ell(\boldsymbol{\xi}; \boldsymbol{\omega}) = \sum_{i=1}^n \ell_i(\boldsymbol{\xi}; \omega_i)$, where, for $\boldsymbol{\omega}_0 = (0, \ldots, 0)^\top$, we have

$$
\begin{aligned}
\ell_i(\boldsymbol{\xi}; \omega_i) = {} & \tau_i(\log(\kappa) + \log(\gamma) + (\kappa - 1)\log(t_i(\omega_i)) + \eta) + \left(1 - \Delta^*(t_i(\omega_i); \boldsymbol{\xi})/\sqrt{\delta + 1}\right) + \\
& + \log\left(\Delta^*(t_i(\omega_i); \boldsymbol{\xi}) + \sqrt{\delta + 1}\right) - \log(2\Delta^*(t_i(\omega_i); \boldsymbol{\xi})) + \\
& + \tau_i \log(\delta(\delta + \Delta(t_i(\omega_i); \boldsymbol{\xi}) + 4\gamma(t_i(\omega_i))^\kappa \exp(\eta) + 3) + 2) + \\
& - 2\tau_i \log(\Delta^*(t_i(\omega_i); \boldsymbol{\xi})(\delta + \Delta(t_i(\omega_i); \boldsymbol{\xi}) + 1)).
\end{aligned}
$$

4.2.3 Covariate perturbation

We consider also an additive perturbation for a specific continuous covariate, X_k say, for $k = 1, \ldots, p$, by setting $x_{ik}(\omega_i) = x_{ik} + \omega_i s_X$, where s_X is a scale factor here assumed to be the standard deviation (SD) of X_k, and $\omega_i \in \mathbb{R}$, for $i = 1, \ldots, n$. Then, the log-likelihood function is $\ell(\boldsymbol{\xi}; \boldsymbol{\omega}) = \sum_{i=1}^n \ell_i(\boldsymbol{\xi}; \omega_i)$,

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

where, for $\boldsymbol{\omega}_0 = (0, \ldots, 0)^\top$ and $\eta_i(\omega_i) = \boldsymbol{x}_i^\top(\omega_i)\boldsymbol{\beta}$,

$$\begin{aligned}\ell_i(\boldsymbol{\xi}; \omega_i) = {} & \tau_i(\log(\kappa) + \log(\gamma) + (\kappa - 1)\log(t_i) + \eta_i(\omega_i)) + 1 - \Delta^*(t_i; \gamma, \kappa, \delta, \eta_i(\omega_i))/\sqrt{\delta + 1} + \\ & + \log\left(\Delta^*(t_i; \gamma, \kappa, \delta, \eta_i(\omega_i)) + \sqrt{\delta + 1}\right) - \log(2\Delta^*(t_i; \gamma, \kappa, \delta, \eta_i(\omega_i))) + \\ & + \tau_i \log(\delta(\delta + \Delta(t_i; \gamma, \kappa, \delta, \eta_i(\omega_i)) + 4\gamma t_i^\kappa \exp(\eta_i(\omega_i)) + 3) + 2) + \\ & - 2\tau_i \log(\Delta^*(t_i; \gamma, \kappa, \delta, \eta_i(\omega_i))(\delta + \Delta(t_i; \gamma, \kappa, \delta, \eta_i(\omega_i)) + 1)).\end{aligned}$$

4.3 Residual analysis

In order to check the goodness of fit of the BS frailty regression model, we propose two types of residuals for this model. These are the generalized Cox–Snell (GCS) and randomized quantile (RQ) residuals proposed by Cox and Snell (1968) and Dunn and Smyth (1996), and given, respectively, by

$$\begin{aligned} r_i^{\mathrm{GCS}} &= -\log(\widehat{S}_T(t_i; \boldsymbol{x}, \boldsymbol{\xi})), \\ r_i^{\mathrm{RQ}} &= \Phi^{-1}(\widehat{S}_T(t_i; \boldsymbol{x}, \boldsymbol{\xi})), \quad i = 1, \ldots, n, \end{aligned} \tag{16}$$

where Φ^{-1} is the inverse function of the N(0, 1) CDF and $\widehat{S}_T(t_i; \boldsymbol{x})$ is the estimated SF and evaluated at the lifetime t_i, that is,

$$\widehat{S}_T(t_i; \boldsymbol{x}, \boldsymbol{\xi}) = \frac{\exp\left(\frac{\widehat{\delta}}{2}\left(1 - \widehat{\Delta}(t_i; \boldsymbol{\xi})\sqrt{\widehat{\delta} + 1}\right)\right)\left(\widehat{\Delta}(t_i; \boldsymbol{\xi}) + \sqrt{\widehat{\delta} + 1}\right)}{2\widehat{\Delta}(t_i; \boldsymbol{\xi})},$$

with

$$\widehat{\Delta}(t_i; \boldsymbol{\xi}) = \sqrt{\widehat{\delta} + 4\widehat{\gamma} t_i^{\widehat{\kappa}} \exp(\widehat{\eta}) + 1}.$$

If the frailty model is correctly specified, then the GCS residual has an EXP(1) distribution, regardless of the frailty model specification, whereas the RQ residual has a N(0, 1) distribution.

5 Applications to medical data sets

In this section, we summarize the proposed methodology by an algorithm and then apply it to two real-world medical data sets. The first one (uncensored) comes from a leukemia study introduced by Feigl and Zelen (1965), whereas the second one (censored) is from a lung cancer trial presented in Kalbfleisch and Prentice (2002). We compare the proposed BS frailty regression model, in terms of fitting, with the Weibull regression model and GA and IG frailty regression models, both of them having a Weibull baseline HR. To make sure that the GA and IG models are identifiable, we consider $U \sim \mathrm{Gamma}(1/\zeta, 1/\zeta)$ and $U \sim \mathrm{IG}(1, \sigma^2)$; see Wienke (2011).

5.1 Summary of the proposed methodology

The proposed methodology is summarized by Algorithm 2.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Algorithm 2. Methodology based on a frailty regression model.

1: Collect n data of a response (usually lifetime), $t_1, \ldots, t_n$ say, and the values of p covariates $(x_{1i}, \ldots, x_{pi})$ for the patient i associated with this response. Data can be censored or not.
2: Carry out an exploratory data analysis for identifying possible candidate models to be considered.
3: Propose a suitable frailty distribution to capture covariates that cannot be observed or measured.
4: Formulate a frailty regression model to estimate the survival probability of a patient according to the general model defined in (3). The formulated frailty regression model must include the response and observed covariates under analysis, as well as the frailty term by its corresponding distribution.
5: Estimate the parameters of the frailty regression model defined in Step 4 and assess the statistical significance of these parameters, as well as the presence of frailty or not by evaluating its variance.
6: Check the frailty regression model estimated in Step 5 by using quantile versus quantile (QQ) plots and GCS and RQ residuals.
7: Compare the model checking in Step 6 to other models (with frailty or not, with covariates or not, nested or not) by using the Akaike (AIC) and Bayesian (BIC) information criteria.
8: Select the best model that describes the data among the compared models in Step 7.
9: Conduct global and local diagnostic studies for the best model selected in Step 8 that describes the data to identify possible influential cases. If no influential cases are detected,
9.1: Then consider as final model to estimate the survival probability of a patient the model selected in Step 8;
9.2: Else, compute the relative change (RC) in the ML estimates of the model parameters and evaluate whether inferential changes are produced or not, when potentially influential cases are removed. If no inferential changes are detected,
9.2.1: Then consider as final model to estimate the survival probability of a patient the model selected in Step 8;
9.2.2: Else, remove the influential cases or propose a robust procedure to estimate the model parameters.

5.2 Application 1: leukemia data

The data set corresponds to the survival times of 33 patients, who died from acute myelogenous leukemia (a kind of cancer that often starts in the bone marrow). Measurements of the patients about white blood cell count at the time of diagnosis were also recorded; see Feigl and Zelen (1965). The patients were separated into two groups depending on the presence or absence of a morphological characteristic of white blood cells. At the time of diagnosis, those patients with the presence of significant granulation of the leukemic cells in the bone marrow were termed as AG positive. The following variables were associated with each studied patient, for $i = 1, \ldots, 33$: (i) T_i is the time from diagnosis to death (in weeks); (ii) X_{i1} is the logarithm of the white blood cell count at the time of diagnosis; and (iii) X_{i2} is the group to which they belong (1: presence—Group 1—or 0: absence—Group 2—of a morphological characteristic).

5.2.1 Exploratory data analysis

Table 1 provides a descriptive summary of the observed lifetime (in weeks) that includes median (MD),

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1 Descriptive statistics for leukemia data.

$t_{(1)}$	MD	$\bar{t}$	SD	CV	CS	CK	$t_{(n)}$	n
1.00	22.00	40.88	46.70	1.14	1.16	3.12	156	33

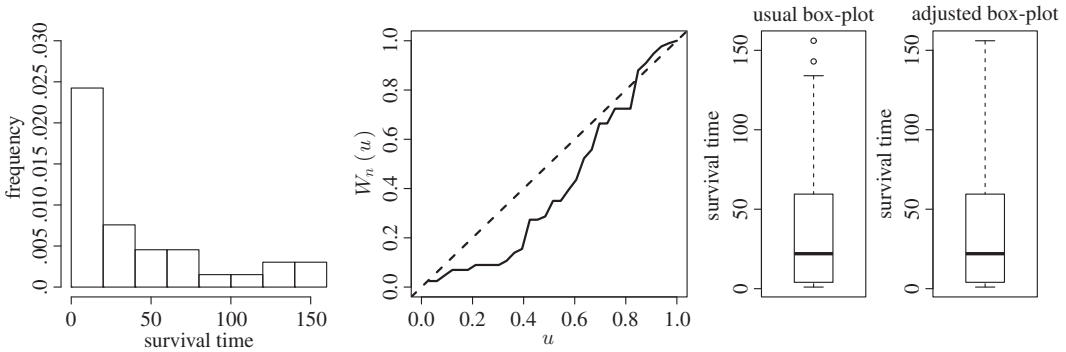

Figure 3 Histogram (left), TTT plot (center), and box-plots (right) for leukemia data.

mean ($\bar{t}$), SD, coefficients of variation (CV), skewness (CS) and kurtosis (CK), and minimum ($t_{(1)}$) and maximum ($t_{(n)}$) values. From this table, note the positively skewed nature and moderate kurtosis level of the data distribution. The skewed nature is confirmed by the histogram of Fig. 3 (left).

The shape of an HR is an important point to decide whether a particular distribution is suitable or not for a data set. A manner to characterize the shape of an HR is by means of the scaled total time on test (TTT) function. We can detect the type of HR that the data have and then choose a suitable distribution. Let $h_T(t) = f_T(t)/(1 - F_T(t))$ be the HR of an RV T, where f_T and F_T are the PDF and CDF of T, respectively. Then, the TTT function is $W(v) = H^{-1}(v)/H^{-1}(1)$, for $0 \leq v \leq 1$, where $H^{-1}(v) = \int_0^{F_T^{-1}(v)} (1 - F_T(z))\mathrm{d}z$, with F_T^{-1} denoting the inverse function of the CDF of T. A plot of the points $(k/n, W_n(k/n))$ can approximate W, with $W_n(k/n) = (\sum_{i=1}^{k} t_{(i)} + (n-k)t_k)/\sum_{i=1}^{n} t_{(i)}$, for $k = 1, \ldots, n$, and $t_{(i)}$ denoting the i-th observed order statistic; see, for example, Fig. 1 in Azevedo et al. (2012) for different theoretical shapes for the scaled TTT curves. Figure 3 (right) suggests a decreasing HR for the observed lifetimes. Therefore, the Weibull distribution is a good choice as baseline HR, since it allows us to model constant, increasing, and decreasing HR, as mentioned. Moreover, this distribution is one of the most used models in survival and reliability analysis due to its good properties and flexibility in data modeling.

Figure 3 (right) presents the usual and adjusted box-plots. The latter box-plot is important in cases where the data follow a skewed distribution, since a significant number of cases can be classified as atypical when they are not; see an `R` package named `robustbase` and use its command `adjbox` to construct the adjusted box-plot; see Rousseeuw et al. (2016). From Fig. 3 (right), note that potential outliers considered by the usual box-plot are not outliers when its adjusted version is considered.

5.2.2 Estimation and model checking

We consider the BS frailty regression model with the structure: $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$. Table 2 provides the estimation and hypothesis testing results for the BS frailty regression model analyzing leukemia data. Results of the Weibull regression model without frailty and the GA and IG frailty regression models also are detailed in this table, as well as their AIC and BIC values. It is worth to highlight

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 2 ML estimates (with estimated asymptotic SEs in parentheses) and model selection measures for the fit to leukemia data with Weibull baseline HR, and respective *p*-values in brackets.

Parameter	Weibull	BS frailty	GA frailty	IG frailty
β_0	−1.903 (1.314)	−17.269 (3.673)	−16.793 (4.711)	−10.978 (2.991)
β_1	1.158 (0.350)	2.975 (0.775)	3.245 (1.047)	1.945 (0.650)
p-value	[<0.0001]	[0.0001]	[0.0020]	[0.0028]
β_2	−0.956 (0.316)	−1.829 (1.050)	−2.108 (0.827)	−1.900 (0.680)
p-value	[<0.0001]	[0.0815]	[0.0108]	[0.0052]
κ	1.001 (0.139)	3.587 (0.695)	1.743 (0.421)	1.704 (0.292)
δ	–	0.015 (0.014)	–	–
ζ	–	–	1.392 (0.827)	–
σ^2	–	–	–	8.312 (8.438)
Log-likelihood	−145.100	−142.109	−144.718	−144.373
AIC	298.144	294.219	299.437	298.747
BIC	304.130	301.701	306.920	306.229

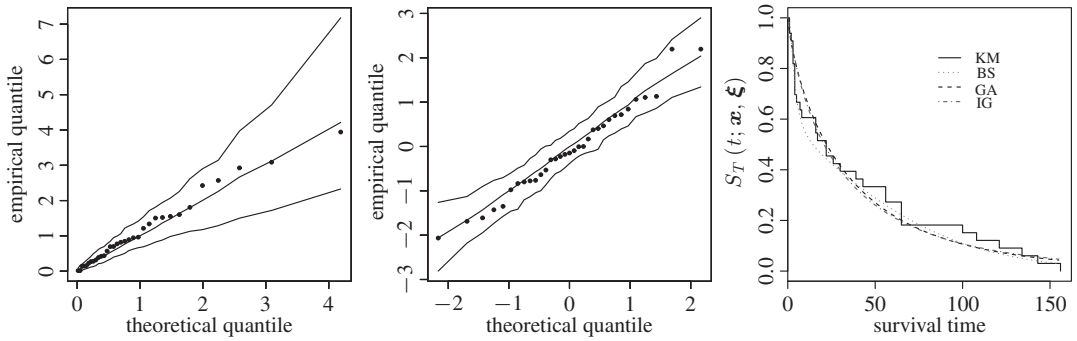

Figure 4 QQ-plot with envelope for the GCS-BS (left) and RQ-BS (center) residuals and fitted SFs (right) with leukemia data.

that, when the models are not nested, such as our case, the AIC and/or BIC should be used to make a decision for the best-fitting model; see Wienke (2011). From Table 2, observe that the BS frailty regression model provides a better fit compared to the other models based on the values of AIC and BIC. Note that, for $\widehat{\delta} = 0.015 < 0.5$, a look at the log-BS distribution reveals bimodality, a behavior not captured by the other models; see Section 2.3. This confirms the flexibility of the proposed model. Figure 4 shows QQ-plots with simulated envelopes for both GCS and RQ residuals defined in (16) based on the BS frailty regression model. Also, this figure shows the fitted SFs based on the Kaplan–Meier (KM) estimator and the BS, GA, and IG frailty models without covariates. The plot of the SFs permits us to compare the empirical and fitted SFs of the data. Figure 4 indicates that, in general, the GCS and RQ residuals present a good agreement with the EXP and N(0, 1) distributions, respectively. Moreover, the fitted SFs confirms graphically the good fit of the BS frailty regression model.

The estimated variance of the BS, GA, and IG frailty regression models are, respectively,

$$\widehat{\mathrm{Var}}(U) = 2\widehat{\delta} + 5/(\widehat{\delta}+1)^2, \quad \widehat{\mathrm{Var}}(U) = \widehat{\zeta}, \quad \widehat{\mathrm{Var}}(U) = \widehat{\sigma}^2. \tag{17}$$

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

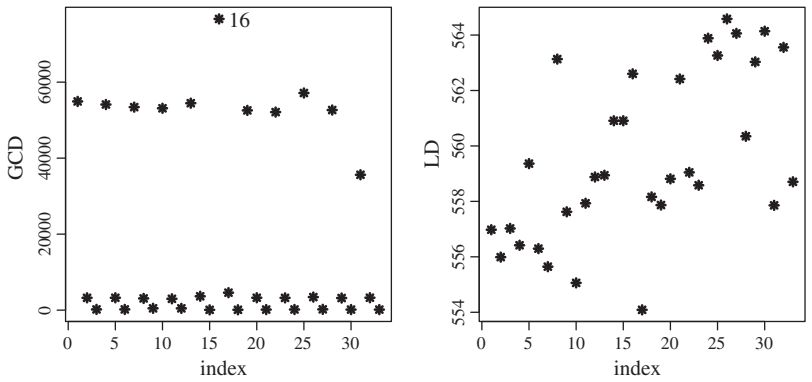

Figure 5 Generalized Cook (left) and likelihood (right) distances for leukemia data.

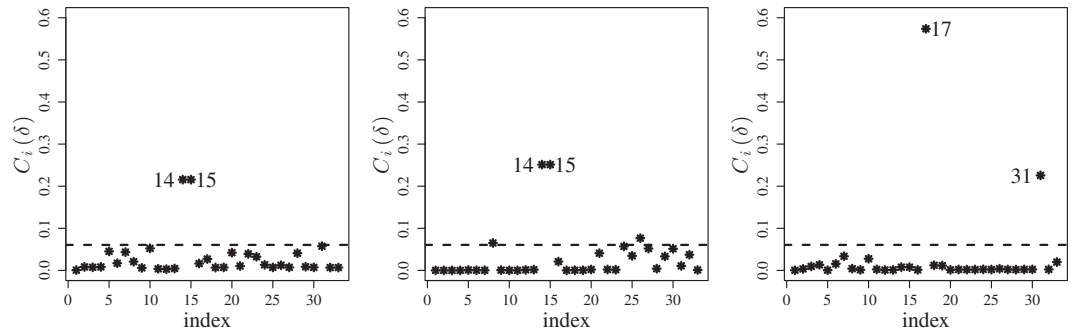

Figure 6 Index plots of C_i for δ under the case-weight (left), response (center), and covariate (right) perturbation schemes with leukemia data.

Note that the variance in the BS case increases when δ is close to zero, that is, a quite small value of δ indicates the presence of high unobserved heterogeneity. Based on Table 2 and expressions given in (17), we compute the corresponding frailty variances. The estimated frailty variances for the BS, GA, and IG frailty regression models based on the leukemia data are 4.881, 1.392, and 8.312, respectively. This indicates the presence of unobserved heterogeneity. Thus, according to the estimated frailty variances, we conclude that the frailty models considered in this study capture the unobserved heterogeneity in the data.

5.2.3 Diagnostic analysis

Next, we carry out our diagnostic analysis based on global and local influence. First, Fig. 5 presents the case-deletion measures $\mathrm{GCD}_i(\boldsymbol{\xi})$ and $\mathrm{LD}_i(\boldsymbol{\xi})$ discussed in Section 4.1. From this figure, on the one hand, the $\mathrm{GCD}_i(\boldsymbol{\xi})$ statistic indicates that the case #16 ($t_{16} = 5.0$, $x_{16,1} = 4.716$, $x_{16,2} = 1.0$) is potentially influential. Note that this case corresponds to a patient with a lifetime of five days (within the lowest values), a logarithm of the number of white blood cells of 4.716 and that belongs to the Group 1, that is, with presence of a morphological characteristic. On the other hand, $\mathrm{LD}_i(\boldsymbol{\xi})$ statistic does not suggest any case as potentially influential.

Index plots of C_i for δ under the case-weight, response, and covariate perturbation schemes are displayed in Fig. 6 (plots corresponding to γ, κ, and $\boldsymbol{\beta}$ look very similar to that for δ and then they are omitted here). Note that the cases #14 and #15 are detected as potentially influential on $\widehat{\delta}$, $\widehat{\kappa}$, and $\widehat{\boldsymbol{\beta}}$ under both the case-weight and response perturbation schemes. The cases #14 and #15 correspond to

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 3 RCs (in %) in ML estimates and their corresponding SEs for the indicated parameter and dropped cases, and respective *p*-values in brackets with leukemia data.

Dropped case		$\widehat{\delta}$	$\widehat{\kappa}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
{14}	RC($\xi_{j(i)}$)	29.16	13.07	0.11	5.71	49.95
	RC(SE($\xi_{j(i)}$))	(28.08)	(11.12)	(4.31)	(3.14)	(2.22)
	p-value	–	–		[<0.0001]	[<0.0001]
{15}	RC($\xi_{j(i)}$)	29.16	13.07	0.11	5.71	49.95
	RC(SE($\xi_{j(i)}$))	(28.08)	(11.12)	(4.31)	(3.14)	(2.22)
	p-value	–	–		[<0.0001]	[<0.0001]
{17}	RC($\xi_{j(i)}$)	4.02	2.22	25.43	33.80	20.10
	RC(SE($\xi_{j(i)}$))	(6.62)	(2.52)	(25.30)	(25.38)	(6.79)
	p-value	–	–		[<0.0001]	[<0.0001]
{31}	RC($\xi_{j(i)}$)	2038.38	41.40	0.49	13.28	57.08
	RC(SE($\xi_{j(i)}$))	(2756.32)	(12.01)	(37.94)	(44.12)	(6.13)
	p-value	–	–		[0.0604]	[0.0093]
{14,15}	RC($\xi_{j(i)}$)	16.49	3.02	2.06	1.70	1.65
	RC(SE($\xi_{j(i)}$))	(18.16)	(0.36)	(0.25)	(0.11)	(0.15)
	p-value	–	–		[<0.0001]	[0.0085]
{17,31}	RC($\xi_{j(i)}$)	378.91	3.77	106.86	153.38	110.68
	RC(SE($\xi_{j(i)}$))	(335.89)	(14.42)	(127.10)	(138.62)	(18.04)
	p-value	–	–		[<0.0001]	[0.0001723]
{14,15,17,31}	RC($\xi_{j(i)}$)	12.44	22.59	42.75	49.96	133.95
	RC(SE($\xi_{j(i)}$))	(53.84)	(47.18)	(75.42)	(94.84)	(16.00)
	p-value	–	–		[0.0045]	[0.0005]

the minimum values of the lifetime of patients (one week in both cases) and the maxima values of the logarithm of the number of white blood cells ($x_{14,1} = 5.0$ and $x_{15,1} = 5.0$). In addition, both of them are in the Group 1, that is, with presence of a morphological characteristic. When the perturbation of the covariate X_{i1} is analyzed, observe that the case #17 ($t_{17} = 65.0$, $x_{17,1} = 5.0$; belonging to the Group 1) and the case #31 ($t_{31} = 30.0$, $x_{31,1} = 4.898$; belonging to the Group 2) are detected as potentially influential on $\widehat{\delta}$, $\widehat{\kappa}$, and $\widehat{\boldsymbol{\beta}}$. It is important to stress that $t_{14} = 1.0$ and $t_{15} = 1.0$ represent the minimum lifetimes of the patients observed in the study and they also present the maximum values for the logarithm of the number of white blood cells ($x_{14,1} = x_{15,1} = 5.0$). In addition, the lifetimes $t_{17} = 65.0$ and $t_{31} = 30.0$ are the maxima values for the logarithm of the number of white blood cells ($x_{17,1} = 5.000$, $x_{31,1} = 4.897$).

In order to check the impact on the model inference of the detected influential cases, we implement the RC. It is computed by removing influential cases and reestimating the parameters as well as their corresponding SEs through the expressions

$$\mathrm{RC}(\xi_{j(i)}) = \left|\frac{\widehat{\xi}_j - \widehat{\xi}_{j(i)}}{\widehat{\xi}_j}\right| \times 100\%, \quad \mathrm{RC}(\mathrm{SE}(\xi_{j(i)})) = \left|\frac{\widehat{\mathrm{SE}}(\widehat{\xi}_j) - \widehat{\mathrm{SE}}(\widehat{\xi}_j)_{(i)}}{\widehat{\mathrm{SE}}(\widehat{\xi}_j)}\right| \times 100\%,$$

where $\widehat{\xi}_{j(i)}$ and $\widehat{\mathrm{SE}}(\widehat{\xi}_j)$ are the ML estimate of ξ_j and its corresponding SE, respectively, after dropping the case i, for $j = 1, \ldots, 5$ and $i = 1, \ldots, 33$, with $\xi_1 = \delta, \xi_2 = \kappa, \xi_3 = \beta_0, \xi_4 = \beta_1$, and $\xi_5 = \beta_2$. Table 3 shows the RCs in the parameter estimates and their corresponding estimated SEs. In addition, *p*-values are shown for the regression coefficients based on t-tests. From this table, note that the largest RCs

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 4 Descriptive statistics for lung cancer data.

$t_{(1)}$	MD	$\bar{t}$	SD	CV	CS	CK	$t_{(n)}$	n
1.00	80.00	121.60	157.82	1.40	3.13	15.55	999.00	137

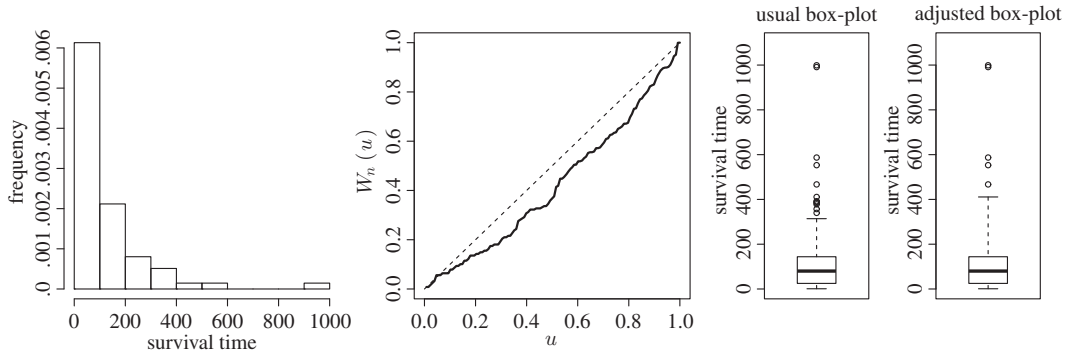

Figure 7 Histogram (left), TTT plot (center), and box-plots (right) for lung cancer data.

are associated with the cases #14, #15, and #31. Observe also that the significance of the parameter estimate of β_2 is altered after removing the cases #14 and #15.

5.3 Application 2: lung cancer data

The data set corresponds to the survival times on 137 advanced lung cancer patients from a Veterans' Administration Lung Cancer trial; see Kalbfleisch and Prentice (2002). The percentage of censored cases is 6.57%. The following variables were associated with each studied patient, for $i = 1, \ldots, 137$: (i) T_i is the lifetime (in days); (ii) X_{i1} is the Karnofsky performance score (where 10–30 is completely hospitalized, 40–60 is partial confinement, 70–90 is able to take care of self, and 100 is a good status of health); and (iii) the tumor-type factor, that is, (X_{i2}) type-1 cell (1 = squamous, 0 = other), (X_{i3}) type-2 cell (1 = small, 0 = other), (X_{i4}) type-3 cell (1 = adeno, 0 = other), and type-4 cell (1 = large, 0 = other).

5.3.1 Exploratory data analysis

Table 4 reports the descriptive statistics of the observed lifetimes (in days) from the mentioned lung cancer trial. The CK and CS indicate the positively skewed nature and high kurtosis level of the data distribution. Figure 7 shows the TTT plot, histogram, and box-plots for these data. Note that the skewed nature is confirmed by the histogram of Fig. 7 (left). The TTT plot displayed in Fig. 7 (center) suggests a decreasing HR for the observed lifetimes, which again justifies the use of the Weibull distribution as a baseline HR. From Fig. 7 (right), notice that some outliers considered by the usual box-plot are not outliers when its adjusted version is considered.

5.3.2 Estimation and model checking

In this case, the regression structure of the model is

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}.$$

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 5 ML estimates (with estimated asymptotic SEs in parentheses) and model selection measures for the fit to lung cancer data, and respective *p*-values in brackets.

Parameter	Weibull	BS frailty	GA frailty	IG frailty
β_0	−1.329 (0.340)	−4.109 (0.616)	−3.625 (0.630)	−4.066 (0.606)
β_1	−0.031 (0.005)	−0.049 (0.011)	−0.057 (0.013)	−0.046 (0.009)
p-value	[<0.0001]	[<0.0001]	[<0.0001]	[<0.0001]
β_2	0.755 (0.246)	1.016 (0.359)	0.613 (0.411)	0.995 (0.340)
p-value	[0.0020]	[0.0046]	[0.1364]	[0.0034]
β_3	1.182 (0.285)	1.363 (0.416)	1.088 (0.435)	1.353 (0.394)
p-value	[<0.0001]	[0.0011]	[0.0123]	[0.0006]
β_4	0.343 (0.269)	0.244 (0.405)	−0.470 (0.527)	0.275 (0.371)
p-value	[0.2010]	[0.5476]	[0.3724]	[0.4580]
κ	1.066 (0.066)	1.469 (0.228)	1.574 (0.243)	1.406 (0.178)
δ		2.043 (1.643)	–	–
ζ		–	0.886 (0.423)	–
σ^2		–	–	0.936 (0.767)
Log-likelihood	−716.510	−712.930	−713.744	−713.248
AIC	1445.030	1439.861	1441.489	1440.497
BIC	1462.550	1460.301	1461.929	1460.937

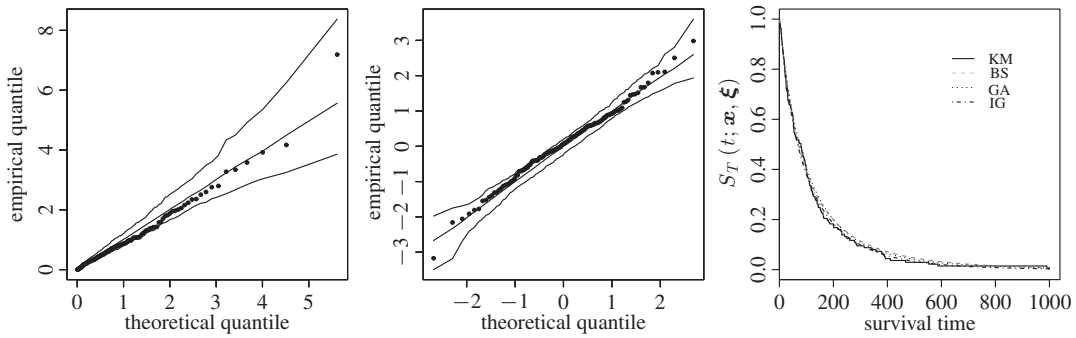

Figure 8 QQ-plot with envelope for the GCS-BS (left) and RQ-BS (center) residuals and fitted SFs (right) with lung cancer data.

The ML estimates of the model parameters, AICs and BICs are reported in Table 5. The results of the information criteria indicate that the BS frailty regression model has the smallest AIC and BIC values, suggesting that it provides the best fit to this data set.

The estimated frailty variances for the indicated model based on lung cancer data are 0.981, 0.886, and 0.936 for the BS, GA, and IG frailty regression models, respectively. This indicates the presence of unobserved heterogeneity. Notice that a slight difference between estimated frailty variances is detected, being it in the BS model slightly greater than in the GA and IG models, indicating that the BS model captures the unobserved heterogeneity in the data in a better way. Figure 8 displays the QQ-plots with simulated envelopes for the GCS and RQ residuals and the fitted SFs based on the KM estimator, as well as on the BS, GA, and IG frailty models without covariates. These graphical plots show the notorious agreement, in terms of fitting to the data, of the BS frailty regression model.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

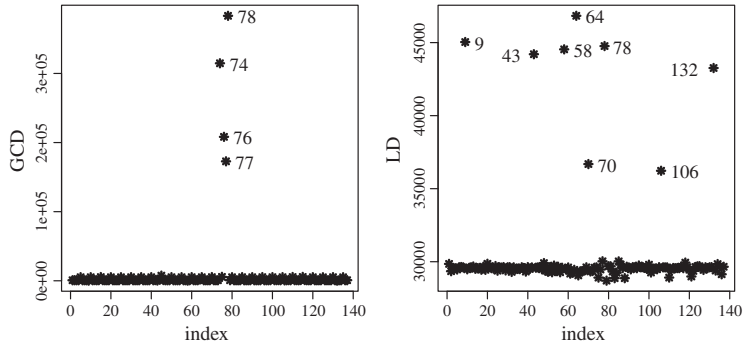

Figure 9 Generalized Cook (left) and likelihood (right) distances for lung cancer data.

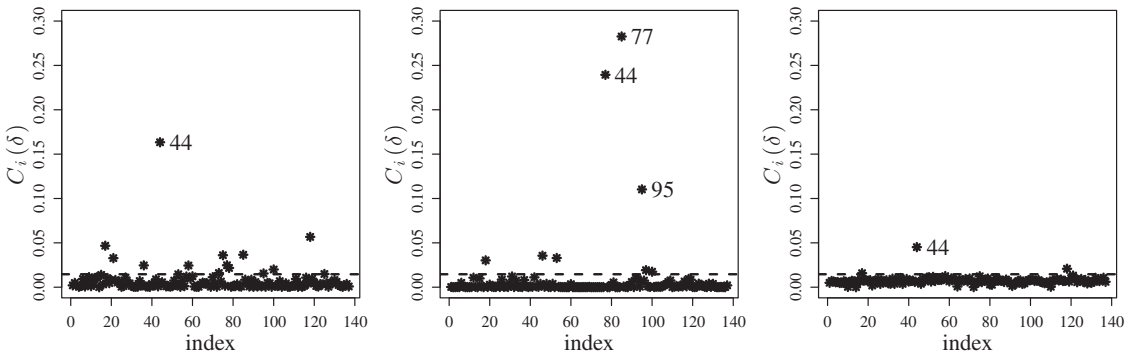

Figure 10 Index plots of C_i for δ under the case-weight (left), response (center), and covariate (right) perturbation schemes with lung cancer data.

We now carry out our diagnostic analysis based on global and local influence. First, Fig. 9 presents the case-deletion measures $\mathrm{GCD}_i(\boldsymbol{\xi})$ and $\mathrm{LD}_i(\boldsymbol{\xi})$. From this figure, note that the $\mathrm{GCD}_i(\boldsymbol{\xi})$ statistic indicates that the cases #74, #76, #77, and #78 are potentially influential. All these cases have type-1 cell and their lifetimes are equal to 242, 111, 1, and 587 days, respectively. Notice that $t_{77} = 1.0$ (one day), that is, it corresponds to the minimum lifetime observed value in the data set. The values of Karnofsky performance score from these cases are 50, 70, 20, and 60, respectively. From Fig. 9 and the $\mathrm{LD}_i(\boldsymbol{\xi})$ statistic, observe that the cases #9, #43, #58, #64, #70, #106, and #132 are potentially influential. The cases #9 and #70 have type-1 cell, whereas the cases #43 and #106 have type-2 cell, and the other cases have type-4 cell. From these cases, the values of the Karnofsky performance score are 50, 70, 90, and 30, respectively, where $x_{9,1} = x_{43,1} = 50$, $x_{64,1} = x_{70,1} = 90$, and $x_{106,1} = x_{132,1} = 30$. In addition, the case #78 ($t_{78} = 587$) is the maximum lifetime, whereas the case #77 ($t_{77} = 1$) is the minimum lifetime. The cases #64 as #70 have the largest values related to the Karnofsky performance score and the cases #77, #106, and #132 have the minimum value of this covariate.

Index plots of C_i for δ under the case-weight, response, and covariate perturbation schemes are displayed in Fig. 10 (such as in Application 1, plots corresponding to γ, κ, and $\boldsymbol{\beta}$ look very similar to that for δ and then they are omitted here). Observe that the case #44 ($t_{44} = 392$, $x_{44,1} = 40$, $x_{44,2} = 0$, $x_{44,3} = 1$, $x_{44,4} = 0$) is detected as potentially influential on $\widehat{\delta}$, $\widehat{\kappa}$, and $\widehat{\boldsymbol{\beta}}$ under the case-weight, response, and covariate perturbation schemes. This case corresponds to a patient with the maximum lifetime (392 days), with almost half of the value considered a good Karnofsky score and small tumor. Regarding the response perturbation, note that the cases #77 ($t_{77} = 1.0$, $x_{77,1} = 20$, $x_{77,2} = 1$, $x_{77,3} = 0$, $x_{77,4} = 0$)

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

and #95 ($t_{95} = 2.0$, $x_{95,1} = 40$, $x_{95,2} = 0$, $x_{95,3} = 1$, $x_{95,4} = 0$) are also detected as potentially influential on $\widehat{\delta}$, $\widehat{\kappa}$, and $\widehat{\boldsymbol{\beta}}$. The cases #77 and #95 correspond to patients with the smallest lifetimes and present squamous and small tumors, respectively. However, the case #77 has a value of Karnofsky score situated within the smallest ones, whereas the case #95 has a value of Karnofsky score around the median value.

Table 6 shows the RCs in the parameter estimates and their corresponding estimated SEs. Also, *p*-values are shown for the regression coefficients based on t-tests. From this table, note that the largest RCs are associated with the set of cases {#44, #95} and {#44, #77, #95}. Observe also that the significance of the parameter β_2 is altered after removing those cases.

6 Discussion, conclusions, and future research

In this paper, we introduced a methodology based on a new regression model with BS distribution for its frailty. In this methodology, a frailty parameter as well as covariates are included in the model to bring further information that may be useful in practice. The inclusion of covariates aims to account for differences in risk, whereas the inclusion of a frailty helps to capture unobserved heterogeneity that covariates may fail to fully account for. This may be due to a missing covariate in the model that can be explained by the frailty. The introduced methodology encompassed inference about the model parameters and influence diagnostics. Also, we considered two types of residuals for the new frailty regression model. The methodology was summarized by an algorithm that allows a practitioner to understand it in a better form. We applied the proposed model to two real-world data sets concerning the survival times of patients who (i) died due to acute myelogenous leukemia or (ii) had advanced lung cancer, considering uncensored and censored data, respectively. We also applied global and local influence diagnostic tools for the proposed model with both of these data sets.

The two applications illustrated the potential of the introduced methodology based on the BS frailty regression model. From a medical point of view, it is important to adequately handle biological variation among individuals. In this sense, good parametric frailty regression models should be used more frequently in medical survival analysis. As a simple example of the applicability of the proposed methodology, one can think of medical doctors, researchers, and/or practitioners estimating the survival time of a patient or group of patients in a clinical study. Moreover, the methodology introduced in this paper may be applied in a medical context to find surrogate measures (specific scores, e.g., concerning the activity of daily living) or to detect frail individuals; see Wienke (2011). We implemented all functions developed in this paper in the `R` software. Then, the use of the introduced methodology becomes easier. The `R` codes used in this study, as well as the data sets, are given in the supplementary material available in the website of *Biometrical Journal*. We hope to report an `R` package in the future with the results obtained in our investigations.

As part of a further research, we leave open the following issues. Economou and Caroni (2008) introduced a manner to construct plots that allow us to verify the correct choice of the frailty distribution. This was conducted in the case of proportional hazard models and for exponential family members, which include the gamma and inverse Gaussian models. These diagnostic plots are based on the closure property that holds for the exponential family and says that the distribution among survivors belongs to the same family of distributions; see Hougaard (1984). The BS model does not hold this property. However, as mentioned in Section 2.3, it is closed under reciprocation, which permits us to construct a graphical procedure, based on the TTT function, to assess the correct choice of the frailty distribution in a proportional hazard model; see Athayde (2016). The construction of this graph is beyond the scope of our study, so that we will consider it as part of a future work. In addition, the introduced methodology can be extended into a model with a nonparametric baseline HR. One may also consider a Bayesian approach to estimate the model parameters. Furthermore, it might be of interest to consider the LR method for testing homogeneity of the frailty term; see Economou and Stehlik (2015). The introduced methodology may also be extended to the correlated frailty case, where the nature of the

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 6 RCs (in %) in ML estimates and their corresponding SEs for the indicated parameter and dropped cases, and respective *p*-values in brackets with lung cancer data.

Dropped case		$\widehat{\delta}$	$\widehat{\kappa}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$
{9}	RC($\xi_{j(i)}$)	3.85	0.84	2.37	2.22	7.57	5.25	30.94
	RC(SE($\xi_{j(i)}$))	(16.17)	(9.41)	(2.09)	(1.21)	(1.18)	(3.91)	(9.11)
	p-value	–	–		[<0.0001]	[<0.0001]	[0.0002]	[0.0677]
{43}	RC($\xi_{j(i)}$)	3.57	0.21	0.42	0.49	0.71	0.21	2.04
	RC(SE($\xi_{j(i)}$))	(0.32)	(6.40)	(4.87	(3.94)	(0.57)	(0.33)	(4.05)
	p-value	–	–		[<0.0001]	[0.0005]	[0.0001]	[0.0539]
{44}	RC($\xi_{j(i)}$)	25.17	0.98	2.10	3.11	18.10	3.19	20.06
	RC(SE($\xi_{j(i)}$))	(56.10)	(14.09)	(56.06)	(11.50)	(0.59)	(2.69)	(3.64)
	p-value	–	–		[<0.0001]	[<0.0001]	[<0.0001]	[0.0438]
{58	RC($\xi_{j(i)}$)	7.96	0.80	0.69	1.74	0.29	0.08	42.12
	RC(SE($\xi_{j(i)}$))	(11.05)	(0.49)	(0.20)	(0.76)	(1.24)	(1.44)	(3.90)
	p-value	–	–		[<0.00001]	[0.0004]	[0.0001]	[0.0374]
{64}	RC($\xi_{j(i)}$)	1.48	0.36	0.43	1.26	0.20	0.08	12.00
	RC(SE($\xi_{j(i)}$))	(14.53)	(14.33)	(1.86)	(6.42)	(1.90)	(1.76)	(4.24)
	p-value	–	–		[<0.00001]	[0.0004]	[0.0001]	[0.0482]
{70}	RC($\xi_{j(i)}$)	0.67	0.22	0.39	2.15	4.86	3.97	27.43
	RC(SE($\xi_{j(i)}$))	(20.30)	(16.22)	(0.60)	(12.94)	(2.06)	(1.25)	(2.10)
	p-value	–	–		[<0.0001]	[0.0006]	[0.0001]	[0.0669]
{74}	RC($\xi_{j(i)}$)	5.52	0.86	1.93	2.04	5.63	3.90	24.10
	RC(SE($\xi_{j(i)}$))	(3.81)	(3.63)	(0.53)	(2.68)	(1.52)	(0.73)	(0.35)
	p-value	–	–		[<0.00001]	[0.0007]	[0.0001]	[0.0649]
{76}	RC($\xi_{j(i)}$)	3.84	0.91	0.10	0.77	1.67	1.06	10.53
	RC(SE($\xi_{j(i)}$))	(24.43)	(16.35)	(5.46)	(7.74)	(0.14)	(0.50)	(2.53)
	p-value	–	–		[<0.0001]	[0.0004]	[0.0001]	[0.0495]
{77}	RC($\xi_{j(i)}$)	14.21	5.06	2.56	1.21	12.75	10.39	27.60
	RC(SE($\xi_{j(i)}$))	(15.97)	(7.52)	(10.10)	(1.39)	(4.83)	(6.43)	(3.22)
	p-value	–	–		[<0.0001]	[<0.0001]	[<0.0001]	[0.0441]
{78}	RC($\xi_{j(i)}$)	0.44	0.63	1.82	0.77	9.46	6.75	40.12
	RC(SE($\xi_{j(i)}$))	(10.46)	(9.05)	(1.94)	(4.73)	(0.88)	(1.23)	(0.05)
	p-value	–	–		[<0.0001]	[0.0010]	[0.0002]	[0.0719]
{95}	RC($\xi_{j(i)}$)	11.53	3.80	3.42	2.93	1.30	3.30	1.00
	RC(SE($\xi_{j(i)}$))	(5.39)	(13.08)	(4.30)	(9.09)	(0.07)	(1.09)	(0.30)
	p-value	–	–		[<0.0001]	[<0.0001]	[<0.0001]	[0.0550]
{106}	RC($\xi_{j(i)}$)	4.13	0.43	1.11	1.89	2.01	0.25	1.16
	RC(SE($\xi_{j(i)}$))	(9.74)	(3.80)	(1.22)	(0.21)	(1.65)	(1.98)	(1.08)
	p-value	–	–		[<0.0001]	[0.0004]	[0.0001]	[0.0548]
{132}	RC($\xi_{j(i)}$)	2.31	0.08	0.04	0.23	0.04	0.10	9.90
	RC(SE($\xi_{j(i)}$))	(10.44)	(7.35)	(0.79)	(2.17)	(1.62)	(1.77)	(1.63)
	p-value	–	–		[<0.0001]	[0.0004]	[0.0001]	[0.0502]
{44,77}	RC($\xi_{j(i)}$)	0.04	7.39	1.18	6.49	32.26	14.01	45.48
	RC(SE($\xi_{j(i)}$))	(18.88)	(25.19)	(20.50)	(15.54)	(8.09)	(5.57)	(0.40)
	p-value	–	–		[<0.0001]	[<0.0001]	[<0.0001]	[0.0367]
{44,95}	RC($\xi_{j(i)}$)	7.21	5.25	5.74	6.85	18.01	6.68	18.89
	RC(SE($\xi_{j(i)}$))	(30.76)	(20.78)	(0.94)	(20.13)	(1.63)	(0.11)	(0.70)
	p-value	–	–		[<0.0001]	[<0.0001]	[<0.0001]	[0.0471]
{44,77,95}	RC($\xi_{j(i)}$)	23.38	14.65	16.23	13.05	34.60	20.31	44.09
	RC(SE($\xi_{j(i)}$))	(30.05	(13.88)	(12.52)	(9.74)	(7.91)	(8.25)	(0.03)
	p-value	–	–		[<0.0001]	[<0.0001]	[<0.0001]	[0.0385]

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

heterogeneity and the dependence are explicitly specified, and are of main importance; see Petersen (1998). Finally, the inclusion of multivariate aspects in frailty models, as well as spatial components, can also be considered; see Garcia-Papani et al. (2016) and Marchant et al. (2016a, 2016b). Work on some of these issues is currently in progress and we hope to report some findings in a future paper.

Acknowledgments The authors thank the Editors and reviewers for their constructive comments on an earlier version of this manuscript that resulted in this improved version. This research work was partially supported by FAPESP, CNPq, and CAPES grants, Brazil, and by FONDECYT 1160868 grant, Chile.

Conflict of interest
The authors have declared no conflict of interest.

References

Aalen, O., Borgan, O. and Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer, New York, NY.

Aalen, O. and Tretli, S. (1999). Analysing incidence of testis cancer by means of a frailty model. *Cancer Causes and Control* **10**, 285–292.

Athayde, E. (2016). A characterization of the Birnbaum-Saunders distribution. *REVSTAT Statistical Journal*, page in press available at https://www.ine.pt/revstat/forthcoming_papers.html.

Azevedo, C., Leiva, V., Athayde, E. and Balakrishnan, N. (2012). Shape and change point analyses of the Birnbaum-Saunders-t hazard rate and associated estimation. *Computational Statistics and Data Analysis* **56**, 3887–3897.

Balakrishnan, N. and Peng, Y. (2006). Generalized gamma frailty model. *Statistics in Medicine* **25**, 2797–2816.

Birnbaum, Z. W. and Saunders, S. C. (1969). A new family of life distributions. *Journal of Applied Probability* **6**, 319–327.

Cai, B. (2010). Bayesian semiparametric frailty selection in multivariate event time data. *Biometrical Journal* **52**, 171–185.

Clayton, D. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467–485.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, Boca Raton, FL.

Cook, R. D. (1987). Influence assessment. *Journal of Applied Statistics* **14**, 117–131.

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London, UK.

Cox, D. and Snell, E. (1968). A general definition of residuals. *Journal of the Royal Statistical Society B* **2**, 248–275.

Cox, D. R. (1972). Regression models and life-tables. *Journal of Royal Statistical Society B* **34**, 187–220.

Crow, E. L. and Shimizu, K. (1988). *Lognormal Distributions: Theory and Applications*. Dekker, New York, NY.

Desmond, A. (1985). Stochastic models of failure in random environments. *Canadian Journal of Statistics* **13**, 171–183.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Springer, New York, NY.

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

Economou, P. and Caroni, C. (2008). Closure properties and diagnostic plots for the frailty distribution in proportional hazards models. In: Vonta, F., Nikulin, M.S., Limnios, N. and Huber-Carol, C. (Eds.), *Statistical Models and Methods for Biomedical and Technical Systems*, chapter 4. Birkhäuser, Boston, pp. 43–53.

Economou, P. and Stehlik, M. (2015). On small samples testing for frailty through homogeneity test. *Communications in Statistics: Simulation and Computation* **44**, 40–65.

Efron, B. and Hinkley, D. (1978). Assessing the accuracy of the maximum likelihood estimator: observed vs. expected Fisher information. *Biometrika* **65**, 457–487.

Elbers, C. and Ridder, G. (1982). True and spurious duration dependence: the identifiability of the proportional hazard model. *Review of Economic Studies* **49**, 403–409.

Espinheira, P., Ferrari, S. and Cribari-Neto, F. (2008). Influence diagnostics in beta regression. *Computational Statistics and Data Analysis* **52**, 4417–4431.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics* **30**, 74–99.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* **21**, 826–837.
Garcia-Papani, F., Uribe-Opazo, M., Leiva, V. and Aykroyd, R. (2016). Birnbaum-Saunders spatial modelling and diagnostics applied to agricultural engineering data. *Stochastic Environmental Research and Risk Assessment*, pages in press available at http://dx.doi.org/10.1007/s00477-015-1204-4.
Henderson, R. and Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society B* **61**, 367–379.
Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika* **71**, 75–84.
Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer, New York, NY.
Johnson, N., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, Vol. **2**. Wiley, New York, NY.
Jones, M. C. (2008). On reciprocal symmetry. *Journal of Statistical Planning and Inference* **138**, 3039–3043.
Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York, NY.
Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, NY.
Kotz, S., Leiva, V. and Sanhueza, A. (2010). Two new mixture models related to the inverse Gaussian distribution. *Methodology and Computing in Applied Probability* **12**, 199–212.
Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley, New York, NY.
Leiva, V. (2016). *The Birnbaum-Saunders Distribution*. Academic Press, New York, NY.
Leiva, V., Marchant, C., Ruggeri, F. and Saulo, H. (2015a). A criterion for environmental assessment using Birnbaum-Saunders attribute control charts. *Environmetrics* **26**, 463–476.
Leiva, V., Rojas, E., Galea, M. and Sanhueza, A. (2014a). Diagnostics in Birnbaum-Saunders accelerated life models with an application to fatigue data. *Applied Stochastic Models in Business and Industry* **30**, 115–131.
Leiva, V., Ruggeri, F., Saulo, H. and Vivanco, J. F. (2017). A methodology based on the Birnbaum-Saunders distribution for reliability analysis applied to nano-materials. *Reliability Engineering and System Safety* **157**, 192–201.
Leiva, V., Santos-Neto, M., Cysneiros, F. J. A. and Barros, M. (2014b). Birnbaum-Saunders statistical modelling: a new approach. *Statistical Modelling* **14**, 21–48.
Leiva, V., Saulo, H., Leão, J. and Marchant, C. (2014c). A family of autoregressive conditional duration models applied to financial data. *Computational Statistics and Data Analysis* **79**, 175–191.
Leiva, V., Tejo, M., Guiraud, P., Schmachtenberg, O., Orio, P. and Marmolejo, F. (2015b). Modeling neural activity with cumulative damage distributions. *Biological Cybernetics* **109**, 421–433.
Marchant, C., Leiva, V. and Cysneiros, F. (2016a). A multivariate log-linear model for Birnbaum-Saunders distributions. *IEEE Transactions on Reliability* **65**, 816–827.
Marchant, C., Leiva, V., Cysneiros, F. and Vivanco, J. (2016b). Diagnostics in multivariate generalized Birnbaum-Saunders regression models. *Journal of Applied Statistics* **43**, 2829–2849.
Mazroui, Y., Mathoulin-Pelissier, S., MacGrogan, G., Brouste, V. and Rondeau, G. (2013). Multivariate frailty models for two types of recurrent events with a dependent terminal event: application to breast cancer data. *Biometrical Journal* **55**, 866–884.
Osorio, F., Paula, G. and Galea, M. (2007). Assessment of local influence in elliptical linear models with longitudinal structure. *Computational Statistics and Data Analysis* **51**, 4354–4368.
Paula, G. A., Medeiros, M. J. and Vilca, F. (2009). Influence diagnostics for linear models with first autoregressive elliptical errors. *Statistical and Probability Letters* **79**, 339–346.
Petersen, J. (1998). An additive frailty model for correlated life times. *Biometrics* **54**, 646–661.
R-Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, AT.
Rieck, J. and Nedelman, J. (1991). A log-linear model for the Birnbaum-Saunders distribution. *Technometrics* **3**, 51–60.
Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Maechler, M. (2016). *robustbase: Basic Robust Statistics*. R package version 0.92-6. Available at http://CRAN.R-project.org/package=robustbase
Santos-Neto, M., Cysneiros, F. J. A., Leiva, V. and Ahmed, S. (2012). On new parameterizations of the Birnbaum-Saunders distribution. *Pakistan Journal of Statistics* **28**, 1–26.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Santos-Neto, M., Cysneiros, F. J. A., Leiva, V. and Barros, M. (2014). On new parameterizations of the Birnbaum-Saunders distribution and its moments, estimation and application. *REVSTAT Statistical Journal* **12**, 247–272.

Santos-Neto, M., Cysneiros, F. J. A., Leiva, V. and Barros, M. (2016). Reparameterized Birnbaum-Saunders regression models with varying precision. *Electronic Journal of Statistics* **10**, 2825–2855.

Saulo, H., Leiva, V., Ziegelmann, F. A. and Marchant, C. (2013). A nonparametric method for estimating asymmetric densities based on skewed Birnbaum-Saunders distributions applied to environmental data. *Stochastic Environmental Research and Risk Assessment* **27**, 1479–1491.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York, NY.

Stare, J. and O'Quigley, J. (2004). Fit and frailties in proportional hazards regression. *Biometrical Journal* **46**, 157–164.

Vanegas, L. and Paula, G. (2016a). An extension of log-symmetric regression models: R codes and applications. *Journal of Statistical Simulation and Computation* **86**, 1709–1735.

Vanegas, L. and Paula, G. (2016b). Log-symmetric distributions: statistical properties and parameter estimation. *Brazilian Journal of Probability and Statistics* **30**, 196–220.

Vaupel, J., Manton, K. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454.

Wienke, A. (2011). *Frailty Models in Survival Analysis*. Chapman and Hall, London, UK.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim