Chilean Journal of Statistics
Vol. 5, No. 2, September 2014, 29–44

Survival Analysis
Research Paper

Generalized time-dependent complement log-log model

Eder Angelo Milani*, Carlos Alberto Ribeiro Diniz and Vera L. D. Tomazella

Department of Statistics, Universidade Federal de São Carlos, São Carlos-SP, Brazil

(Received: 15 August 2013 · Accepted in final form: 08 September 2014)

Abstract

In this article we introduce a new family of survival models of non-proportional hazards. As an extension of this family, we present models with gamma and inverse Gaussian frailty distributions and the unconditional survival function is derived by using Laplace transform. A simulation study to check the frequentist properties of the proposed models is presented. The developed methodology is illustrated considering a real data set of lung cancer. The proposed models are compared with two models presented in the literature.

Keywords: Non-proportional hazard · Fragility · Function complementary log-log · Model CLL with gamma frailty · Model CLL with inverse Gaussian frailty.

Mathematics Subject Classification: Primary 62N01 · Secondary 62N99.

1. Introduction

Let T be a random variable representing the lifetime of individuals (or components). The use of equivalent functions, such as the cumulative distribution, density, survival and hazard functions, which uniquely determine the distribution of T, is common in survival analysis. The hazard function is manly used due to its interpretation. It is the instantaneous probability of failure of an individual changeing over time and is statistically expressed as $h(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \leq T < t+\Delta t | t \geq t)}{\Delta t}$. The Cox proportional hazard model, Cox (1972), which is a standard approach to survival data, presents the assumption that ratio of the failure rates of any two individuals are proportional. This assumption is not true in several practical real situations and it is usual to find non-proportional hazard data. From this fact, several types of non-proportional hazard models have been created. Klabfleish and Prentice (2002) present the accelerated failure model, Ciampi and Etezadi-Amoli (1985) present the hybrid hazard model, Louzada-Neto (1997 and 1999) present the extended hybrid hazard model and Mackenzie (1996) presents the generalized time-dependent logistic model.

Functions relating the distribution parameters to explanatory variables are called link functions. Functions logit, probit and complementary log-log are usually used in generalized linear models for location, scale and shape, for more details see McCullagh and Nelder (1898). Mackenzie (1996) proposed a new class of survival models using hazard function

*Eder Angelo Milani. Email: edinhomilani@hotmail.com

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

as a logit function, that is,

$$h(t|\alpha,\boldsymbol{\beta}) = \frac{\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}{1+\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}.$$

This is a non-proportional hazard model called Time-Dependent Logistic (TDL) model.

In this paper, we used hazard function as a complementary log-log function. This function is a non-proportional risk and may assume increasing, decreasing and constant behavior.

The usual survival models can be extended with the inclusion of a frailty term. This frailty term can capture the effect of covariates which are important to explain the survival time of the individuals but for some reason, these covariates were not incorporated in the model. It is possible to include the frailty term in the model in an additive or a multiplicative form. These approaches can be found in Tomazella et al. (2006) and in Milani (2011). In Milani (2011) a model extension of the Time-Dependent Logistic model is presented including a multiplicative form of the frailty term with the latent process following a gamma distribution depending on the frailty parameter θ. In this case, the hazard function unconditional on the individual frailty is given by

$$h(t|\alpha,\boldsymbol{\beta},\theta) = \frac{\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}{\left[1+\frac{\theta}{\alpha}\ln\left(\frac{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)\right](1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta}))}.$$

This model is called Time-Dependent Logistic Frailty model (TDLF).

In this paper we propose a new family of survival models with the assumption of non-proportional hazard functions. This family is extended to include a frailty term, following gamma or inverse guassian distribution, in a multiplicative form in the hazard function. Using Laplace transformation we obtain the survival function unconditional on the individual frailty. A simulation study and an application with real data are shown to illustrate the methodology development.

The paper is organized as follows. Section 2 presents the generalized time-dependent complementary log-log models and generalized time-dependent complementary log-log frailty models. Section 3 presents the construction of the likelihood and estimation procedure for both models. Section 4 presents a simulation study considering the bias, square root of the mean-square error and coverage probabilities of the maximum likelihood estimates. A real dataset on lung cancer in Northern Ireland is analyzed using the proposed model and this fitted model is compared with other non-proportional hazard models in Section 5. The conclusions are in Section 6.

2. Complementary Log-Log Hazard Model

Let T be a non-negative random variable representing the failure time of an individual. The complementary log-log hazard function (CLL) is given by

$$h(t|\alpha,\boldsymbol{\beta}) = \exp(-\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})), \tag{1}$$

where α is a measure of the time effect, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a $p+1-$dimensional vector of regression coefficients associated with fixed covariates $\mathbf{x} = (1, x_1, \ldots, x_p)^T$.

From the hazard equation given in (1) the cumulative hazard function is expressed by

$$H(t|\alpha,\boldsymbol{\beta}) = \int_0^t \exp(-\exp(\alpha y + \mathbf{x}'\boldsymbol{\beta}))dy, \tag{2}$$

and from equation (2) the survival function is given by

$$\begin{aligned} S(t|\alpha,\boldsymbol{\beta}) &= \exp(-H(t|\alpha,\boldsymbol{\beta})) \\ &= \exp\left(-\int_0^t \exp(-\exp(\alpha y + \mathbf{x}'\boldsymbol{\beta}))dy\right), \end{aligned} \tag{3}$$

with $S(0|\alpha,\boldsymbol{\beta}) = 1$ for $\alpha \in \mathbb{R}$. For $\alpha \leq 0$, $\lim_{t\longrightarrow\infty} S(t|\alpha,\boldsymbol{\beta}) = 0$ and for $\alpha > 0$, $\lim_{t\longrightarrow\infty} S(t|\alpha,\boldsymbol{\beta}) > 0$, that is, the survival function is proper for $\alpha \leq 0$ and improper for $\alpha > 0$. Some examples of survival function are shown in Figure 1. The behavior of the hazard function depends on the value of α. For $\alpha = 0$ the hazard function is constant; for $\alpha < 0$, the hazard function increases and for $\alpha > 0$ the hazard function decreases. Figure 1 shows some examples of possible shapes of the hazard function.

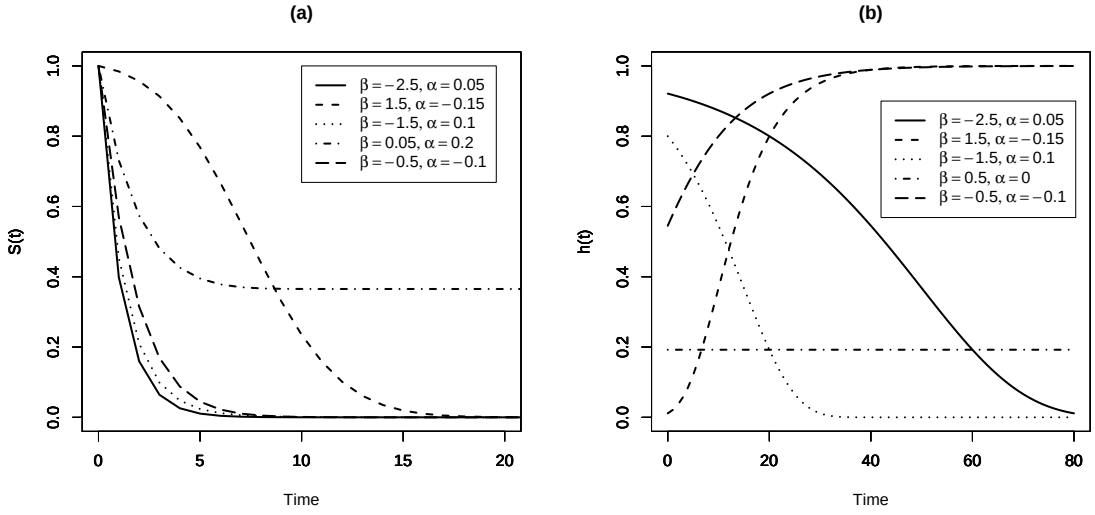

Figure 1. Possible shapes of the survival function in (a) and the hazard function in (b) of the CLL model

The ratio of the hazard function of two individuals is expressed by

$$\begin{aligned} \frac{h(t|\alpha,\boldsymbol{\beta},\mathbf{x_1})}{h(t|\alpha,\boldsymbol{\beta},\mathbf{x_2})} &= \frac{\exp(-\exp(\alpha t + \mathbf{x_1}'\boldsymbol{\beta}))}{\exp(-\exp(\alpha t + \mathbf{x_2}'\boldsymbol{\beta}))} \\ &= \exp[-\exp(\alpha t)(\exp(\mathbf{x_1}'\boldsymbol{\beta}) - \exp(\mathbf{x_2}'\boldsymbol{\beta}))], \end{aligned}$$

where this ratio is constant and equal to 1 if $\mathbf{x_1}'\boldsymbol{\beta} = \mathbf{x_2}'\boldsymbol{\beta}$. Therefore, for individuals where $\mathbf{x_1}'\boldsymbol{\beta} \neq \mathbf{x_2}'\boldsymbol{\beta}$, the time effect does not disappear and, consequently, the non-proportionality is evident.

From equations (1) and (3), the probability density function is given by

$$f(t|\alpha,\boldsymbol{\beta}) = \exp\left(-\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta}) - \int_0^t \exp(-\exp(\alpha y + \mathbf{x}'\boldsymbol{\beta}))dy\right),$$

where $\alpha \leq 0$ and $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$.

2.1 Complementary Log-Log Hazard Frailty Model

Let T_i $(i = 1, \cdots, n)$ be the survival time for ith subject. Denote by V_i the unobserved frailty (or random effect) for the ith subject. We extend the model (1) to include a frailty term acting multiplicatively on the individual hazard rate. Given $V_i = v_i$, the conditional hazard function of T_i takes the form

$$h_i(t|v_i) = v_i \exp(-\exp(\alpha t_i + \mathbf{x}_i'\boldsymbol{\beta})), \tag{4}$$

as v_i represents a value of an unobservable random variable V_i, the individual hazard increases if $v_i > 1$, decreases if $v_i < 1$ and if $v_i = 1$ the frailty model (4) reduces to the complementary log-log (1).

The key idea of this model is that individuals have different frailties, and that the more frail ones will die earlier than the ones who are less frail, hence the name frailty. The frailty term in this model not only explains the heterogeneity among individuals, it also enables us to assess the covariates effect that for some reason were not considered in the planning. For instance, if an important covariate was not included in the model, this will increase the unobservable heterogeneity, affecting the inferences about the parameters related to the covariates in the model. Thus, if we include the frailty term in the model, it will help relieve this problem.

Frailties V_i are assumed to be independent and identically distributed random variables. Due to the way the frailty term acts on the hazard function, natural frailty distribution candidates are supposed to be continuous and time independent, such as gamma, lognormal, Weibull and inverse Gaussian distributions (Hougaard, 1995). In various articles, the gamma distribution is often applied, because it presents an easy algebraic treatment. In this article, V_i follows a gamma or an inverse Gaussian distribution.

Let Z be a random variable following a gamma distribution with parameters τ and η, $\mathrm{G}(\tau, \eta)$, with density function written as

$$f_Z(z) = \frac{\eta^\tau}{\Gamma(\tau)} z^{(\tau-1)} \exp(-z\eta),$$

and let W be a random variable following an inverse Gaussian distribution with parameter ν and λ, $\mathrm{GI}(\nu, \lambda)$, with the density function written as

$$f_W(w) = \lambda^{1/2}(2\pi)^{-1/2} w^{-3/2} \exp\left(-\frac{\lambda}{2\nu^2 w}(w-\nu)^2\right).$$

According to Elbers and Ridder (1982) when working with frailty the random effect distribution must have a finite mean for the model to be identifiable. Taking this into account, in order to keep the identifiability of the model it is convenient to consider the distribution with mean 1. Therefore, we assume a gamma distribution with parameters $\tau = \eta = \theta^{-1}$, where $Var(Z) = \theta$. In the inverse Gaussian distributions, we assume $\nu = 1$ and $\lambda = 1/\sigma^2$, so $Var(W) = \sigma^2$. Note that if $\theta = 0$ or $\sigma^2 = 0$, all frailty variables are equal to 1 , i.e, distributions are degenerated at point 1 and, thus we obtain the model without fragility.

To obtain the likelihood function it is necessary to find the unconditional survival function. The unconditional survival function is given by

$$S(t) = \int_0^\infty S(t|v) g(v) dv, \tag{5}$$

where $g(v)$ is the probability density function of the $\mathrm{G}(1/\theta, 1/\theta)$ or $\mathrm{GI}(1, 1/\sigma^2)$ distribution. To calculate the marginalization, the Laplace transform can be used since both have the same shape.

2.1.1 CLL model with gamma frailty terms (CLLGF)

The Laplace transform of the gamma distribution $\mathrm{G}(1/\theta, 1/\theta)$, is given by

$$Q(s) = (1 + \theta s)^{-1/\theta}, \tag{6}$$

where s is a real argument (for details see Wienke, 2011). Substituting s for $H(t|\alpha, \boldsymbol{\beta})$ in equation (6) we obtain the unconditional survival function, given by,

$$\begin{aligned} S(t|\alpha, \boldsymbol{\beta}, \theta) &= (1 + \theta H(t|\alpha, \boldsymbol{\beta}))^{-1/\theta} \\ &= \left(1 + \theta \int_0^t \exp(-\exp(\alpha y + \mathbf{x}'\boldsymbol{\beta}))dy\right)^{-1/\theta}, \end{aligned} \tag{7}$$

and the correspondent function of the cumulative risk is given by

$$\begin{aligned} H(t|\alpha, \boldsymbol{\beta}, \theta) &= -\ln[(1 + \theta H(t|\alpha, \boldsymbol{\beta}))^{-1/\theta}] \\ &= -\ln\left[\left(1 + \theta \int_0^t \exp(-\exp(\alpha y + \mathbf{x}'\boldsymbol{\beta}))dy\right)^{-1/\theta}\right]. \end{aligned} \tag{8}$$

Deriving the function of the cumulative risk (8) we obtain the correspondent hazard function which is given by

$$h(t|\alpha, \boldsymbol{\beta}, \theta) = \frac{\exp(-\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta}))}{1 + \theta \int_0^t \exp(-\exp(\alpha y + \mathbf{x}'\boldsymbol{\beta}))dy}. \tag{9}$$

The behavior of the hazard function (9) takes several forms, according to the value of α. For $\alpha \geq 0$, the hazard function decreases; for $\alpha < 0$, the hazard function increases or is unimodal, depending on the value of $\mathbf{x}'\boldsymbol{\beta}$. The behaviour of the survival function given in (7) is determined by the value of α. For $\alpha \leq 0$, $S(0|\lambda, \alpha, \boldsymbol{\beta}) = 1$ and $S(\infty|\lambda, \alpha, \boldsymbol{\beta}) = \lim_{t \longrightarrow \infty} S(t|\lambda, \alpha, \boldsymbol{\beta}) = 0$. In other words, the survival function is proper, and for $\alpha > 0$, $S(0|\lambda, \alpha, \boldsymbol{\beta}) = 1$ and $S(\infty|\lambda, \alpha, \boldsymbol{\beta}) \neq 0$, the survival function is improper.

In Figure 2, we illustrate the shapes of the hazard and survival functions of the model CLLGF.

2.1.2 CLL model with inverse gaussian frailty terms (CLLIGF)

The Laplace transform of the inverse gaussian distribution IG $(1, 1/\sigma^2)$ is given by

$$Q(s) = \exp\left[\frac{1}{\sigma^2}\left(1 - \sqrt{2\sigma^2 s + 1}\right)\right], \tag{10}$$

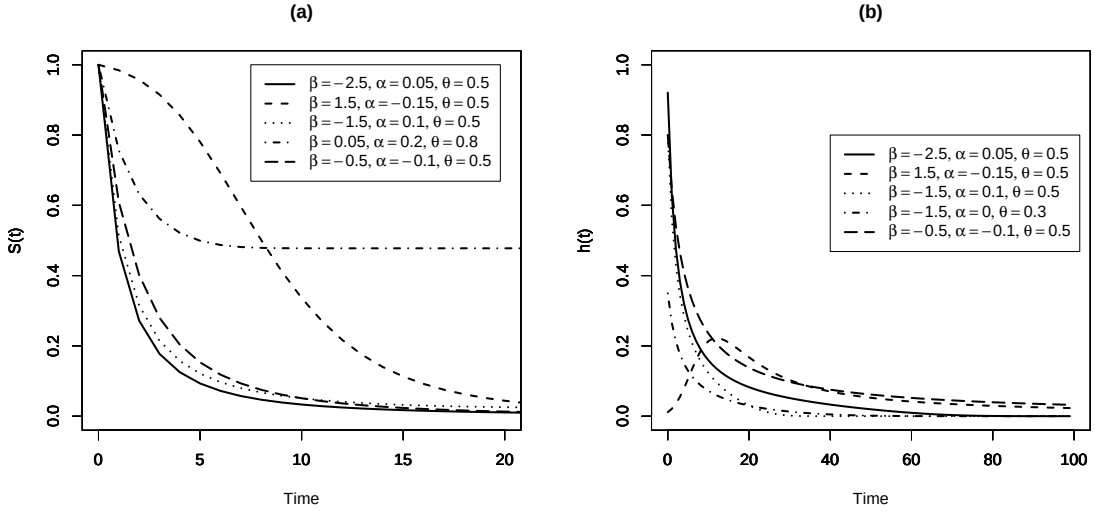

Figure 2. Possible shapes of the survival function in (a) and the hazard function in (b) of the CLLGF model

where s is a real argument. Substituting s for $H(t|\alpha,\boldsymbol{\beta})$ in equation (10), we obtain the unconditional survival function, given by,

$$\begin{aligned} S(t|\alpha,\boldsymbol{\beta},\sigma^2) &= \exp\left[\frac{1}{\sigma^2}\left(1-\sqrt{2\sigma^2 H(t|\alpha,\boldsymbol{\beta})+1}\right)\right] \\ &= \exp\left[\frac{1}{\sigma^2}\left(1-\sqrt{2\sigma^2\int_0^t \exp(-\exp(\alpha y+\mathbf{x}'\boldsymbol{\beta}))dy+1}\right)\right]. \end{aligned} \tag{11}$$

Using (11), the correspondent function of the cumulative risk is given by

$$H(t|\alpha,\boldsymbol{\beta},\sigma^2) = -\frac{1}{\sigma^2}\left(1-\sqrt{2\sigma^2\int_0^t \exp(-\exp(\alpha y+\mathbf{x}'\boldsymbol{\beta}))dy+1}\right). \tag{12}$$

Deriving the function of the cumulative risk (12), we obtain the correspondent hazard function which can be written as

$$\begin{aligned} h(t|\alpha,\boldsymbol{\beta},\sigma^2) &= \frac{dH(t|\alpha,\boldsymbol{\beta},\sigma^2)}{dt} \\ &= \frac{\exp(-\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta}))}{\sqrt{2\sigma^2\int_0^t \exp(-\exp(\alpha y+\mathbf{x}'\boldsymbol{\beta}))dy+1}}. \end{aligned} \tag{13}$$

The behavior of the hazard and survival function are similar to the respective functions in the CLLGF model.

In Figure 3, we illustrate the shapes of the hazard and survival functions of the model CLLIGF.

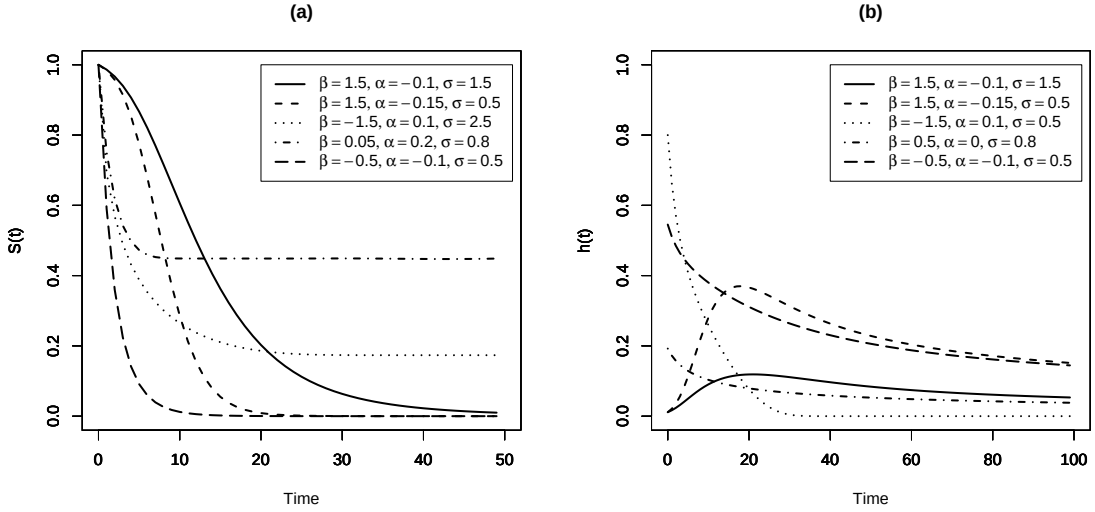

Figure 3. Possible shapes of the survival function in (a) and the hazard function in (b) of the CLLIGF model

3. Likelihood functions

Let $(t_i, \mathbf{x_i}, \delta_i)$, $i = 1, \ldots, n$ be n observed times, $\mathbf{x_i} = (1, x_1, \ldots, x_p)$ a set of observed covariates and δ_i an indicator variable, assuming value 1 for observed failure and 0 for censoring.

The likelihood function for right-censored data may be written as

$$L(\varphi|\mathbf{t}, \mathbf{x}, \boldsymbol{\delta}) = \prod_{i=1}^{n} \{[h(\varphi|\mathbf{t}, \mathbf{x}, \boldsymbol{\delta})]^{\delta_i} S(\varphi|\mathbf{t}, \mathbf{x}, \boldsymbol{\delta})\}.$$

The log-likelihood functions, $l(\varphi|\mathbf{t}, \mathbf{x}, \boldsymbol{\delta}) = \ln[L(\varphi|\mathbf{t}, \mathbf{x}, \boldsymbol{\delta})]$, for the models in (1), (9) and (13) are given by

$$l(\alpha, \boldsymbol{\beta}|\mathbf{t}, \mathbf{x}, \boldsymbol{\delta}) = \sum_{i=1}^{n} \left(- \int_0^{t_i} \exp[-\exp(\alpha y + \mathbf{x_i}'\boldsymbol{\beta})] dy \right) + \sum_{i=1}^{n} -\delta_i \exp(\alpha t_i + \mathbf{x_i}'\boldsymbol{\beta}), \tag{14}$$

$$l(\alpha, \boldsymbol{\beta}, \theta|\mathbf{t}, \mathbf{x}, \boldsymbol{\delta}) = \sum_{i=1}^{n} (-1/\theta - \delta_i) ln \left[1 + \theta \left(\int_0^{t_i} \exp[-\exp(\alpha y + \mathbf{x_i}'\boldsymbol{\beta})] dy \right) \right] + \sum_{i=1}^{n} -\delta_i \exp(\alpha t_i + \mathbf{x_i}'\boldsymbol{\beta}), \tag{15}$$

$$l(\alpha, \boldsymbol{\beta}, \sigma^2|\mathbf{t}, \mathbf{x}, \boldsymbol{\delta}) = \sum_{i=1}^{n} -\delta_i \exp(\alpha t_i + \mathbf{x_i}'\boldsymbol{\beta}) + \sum_{i=1}^{n}(-\delta_i/2)\ln\left[2\sigma^2\left(\int_0^{t_i}\exp[-\exp(\alpha y + \mathbf{x_i}'\boldsymbol{\beta})]dy\right)+1\right] + \sum_{i=1}^{n}(1/\sigma^2)\left[1-\sqrt{2\sigma^2\left(\int_0^{t_i}\exp[-\exp(\alpha y + \mathbf{x_i}'\boldsymbol{\beta})]dy\right)+1}\right]. \quad (16)$$

The maximum likelihood estimators (MLEs) of the parameters of the models in (1), (9) and (13) are obtained, respectively, by direct maximization of the log-likelihood functions (14), (15) and (16), using for instance the L-BFGS-B algorithm (Byrd et al., 1995). The advantage of this procedure is that it runs easily from a statistical package such as R. The asymptotic confidence intervals are obtained by considering the maximum likelihood estimators and the inverse of the Fisher observed information matrix.

4. Simulation Study

This simulation study assesses the square root of the mean squared error (SRMSE) and bias of the MLEs as well as the empirical coverage probabilities of the asymptotic confidence intervals for the parameters of the CLL, CLLGF and CLLIGF models.

The simulated data sets are generated from the following procedure:

(1) Fix values for the model parameters.
(2) Generate a value, u, from a uniform distribution $U[0, 1]$ and a value, v, from the frailty distribution (in case of a frailty model).
(3) Generate a value, x, from a Bernoulli distribution with a known success probability.
(4) Find t_1 from the equation $u = S(t|\text{parameters})$. For the CLL model, the parameter set is (α, $\boldsymbol{\beta}$) and for CLLGF and CLLIGF models, the parameter set is α, $\boldsymbol{\beta}$, θ or σ^2.
(5) Generate a value, t_2, from an exponential distribution and consider $t = min(t_1, t_2)$.
(6) Fix values for δ. $\delta = 1$ if t=t_1 and $\delta = 0$ otherwise.
(7) Repeat steps 2 to 6 until the proper size sample has been acquired.

In step 4, the bisection method (Ruggiero and Lopes, 1997) is used in order to find the root of equation $g(t) = S(t|\text{parameters}) - u$ and the Gauss-Legendre integration formula (Franco, 2010) is used to calculate the value of $S(t|\text{parameters})$. In step 5, the parameter of the exponential distribution is defined in such a way that the proportion of censoring times is around 21%.

The frequentist properties of the maximum likelihood estimators of the parameters are based on 1,000 simulations for each sample size (n = 100, 200, 500 and 1,000). For each simulation we obtain the MLEs and the Hessian matrix. Using these values we calculate the bias, SRMSE and the asymptotic 95% confidence intervals. The empirical coverage probability is calculated as the quotient between the number of intervals containing the true parameter value and the total number of intervals constructed (1,000 intervals).

The values for CLL, CLLGF and CLLIGF model parameters were fixed at $\alpha = -0.5$, $\beta_0 = 1.0$ and $\beta_1 = -0.8$. The covariates were generated from a Bernoulli distribution with a success probability equal to 0.50.

To investigate the effect of the frailty value in the metrics of interest for the CLLGF model, we use $\theta = 0.1, 0.5$ and 0.9; and $\sigma^2 = 0.1, 0.5$ and 0.9 for the CLLIGF model.

Table 1 presents the results for the CLL model. The values of bias and square root of the mean-square error on these Tables are the means of 1,000 values of bias and SRMSE, for each simple sample and for each parameter. The results show that the values are near zero. We also note that the empirical coverage probabilities are close to nominal ones for all sample sizes and for all parameters.

Table 1. Results for CLL model

n	α			β_0			β_1		
	Bias	SRMSE	CP	Bias	SRMSE	CP	Bias	SRMSE	CP
100	0.047	0.195	0.950	0.057	0.277	0.964	0.052	0.293	0.959
200	0.019	0.120	0.943	0.017	0.188	0.953	0.016	0.195	0.957
500	0.008	0.075	0.948	0.007	0.115	0.946	0.005	0.125	0.946
1000	0.005	0.050	0.950	0.006	0.080	0.953	0.005	0.090	0.938

Table 2 present the square root of the mean-square error for CLLGF and CLLIGF models. Note that the values are near zero, regardless of the heterogeneity imposed to the data.

Table 2. SRMSE for CLLGF and CLLIGF models considering different values for θ and σ^2

θ/σ^2		Model CLLGF				Model CLLIGF			
		n							
		100	200	500	1000	100	200	500	1000
0.1									
	α	0.123	0.053	0.014	0.004	0.156	0.075	0.020	0.005
	β_0	0.126	0.052	0.015	0.002	0.152	0.080	0.013	0.008
	β_1	0.105	0.036	0.014	0.003	0.114	0.059	0.010	0.009
	σ^2	0.053	0.022	0.005	0.001	0.109	0.046	0.013	0.004
0.5									
	α	0.075	0.028	0.002	0.002	0.101	0.030	0.008	0.008
	β_0	0.092	0.035	0.009	0.005	0.096	0.033	0.008	0.006
	β_1	0.072	0.027	0.013	0.005	0.060	0.027	0.001	0.003
	σ^2	0.015	0.017	0.013	0.006	0.029	0.001	0.007	0.001
0.9									
	α	0.060	0.017	0.010	0.003	0.109	0.035	0.014	0.004
	β_0	0.063	0.012	0.011	0.005	0.101	0.029	0.012	0.003
	β_1	0.050	0.011	0.014	0.006	0.071	0.007	0.003	0.004
	σ^2	0.033	0.020	0.005	0.002	0.036	0.015	0.005	0.002

Figure 4 shows the behaviour of the square root of the mean-square error for the parameters of the CLLGF model. Note that for different values of θ, the values of SRMSE decrease as the sample sizes increase. Similar results are presented in the CLLIGF model.

Table 3 present the empirical coverage probabilities for each sample size for both models, CLLGF and CLLIGF. Note that the empirical coverages are greater than the nominal ones in the situation where the data presents little heterogeneity and a small sample size (n = 100 and 200). For sample size n = 500 or 1000, the empirical coverage probabilities are near the nominal ones. Considering other scenarios of θ and sample size, the empirical coverage probabilities are all close to the nominal ones.

In order to verify if the CLL model and its extensions present different results, we conducted a simulation study generating data from a model and obtaining a maximum

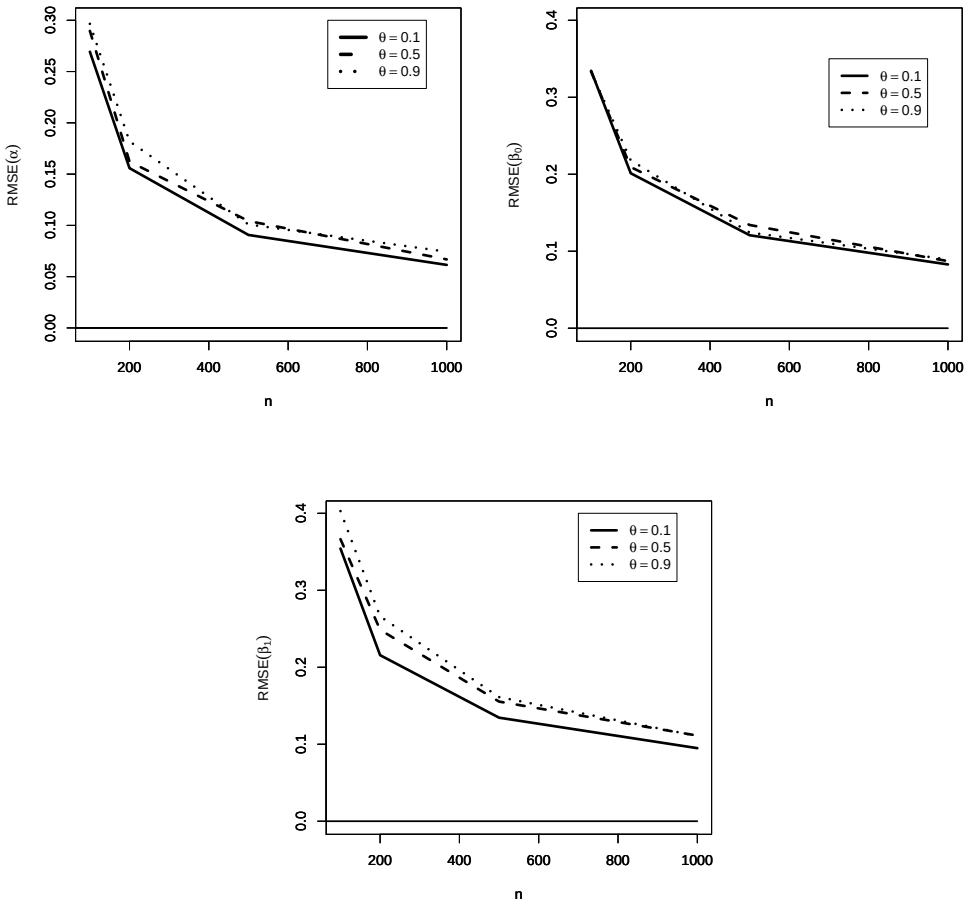

Figure 4. Square root of the mean-square error for the parameters of the CLLGF model

Table 3. Empirical coverage probabilities for CLLGF and CLLIGF models

θ/σ^2		Model CLLGF				Model CLLIGF			
		n							
		100	200	500	1000	100	200	500	1000
0.1									
	α	0.982	0.971	0.961	0.956	0.984	0.967	0.962	0.952
	β_0	0.963	0.960	0.957	0.955	0.972	0.959	0.962	0.965
	β_1	0.969	0.960	0.955	0.948	0.972	0.955	0.964	0.965
	θ/σ^2	0.993	0.990	0.984	0.958	0.996	0.993	0.986	0.976
0.5									
	α	0.946	0.950	0.934	0.954	0.967	0.947	0.946	0.939
	β_0	0.964	0.956	0.938	0.957	0.972	0.955	0.939	0.936
	β_1	0.966	0.947	0.945	0.932	0.974	0.947	0.963	0.940
	θ/σ^2	0.945	0.943	0.947	0.948	0.954	0.936	0.934	0.934
0.9									
	α	0.930	0.925	0.957	0.943	0.946	0.933	0.949	0.948
	β_0	0.949	0.942	0.961	0.950	0.968	0.948	0.948	0.936
	β_1	0.955	0.945	0.953	0.941	0.967	0.949	0.952	0.948
	θ/σ^2	0.938	0.937	0.950	0.954	0.917	0.939	0.936	0.938

likelihood estimative for the parameters of all other models. The TDL and TDLF models are also included in this study. The best model is evaluated using AIC and BIC. This procedure is repeated 1,000 times for each model used in the generation. The number of times which the criteria chooses the correct model is shown in Table 4. We observe that the correct number of time increases as we increase the sample size. The choice made by the criteria is not clear for cases with a moderate sample size, n = 100 or 200.

Table 4. Number of times which the criteria, AIC and BIC, choose the correct model

		Estimative				
Generated from	N	TDL	TDLF	CLL	CLLGF	CLLIGF
CLL						
	100	419/421	38/8	494/562	11/0	38/9
$\alpha = -0.3, \beta_0 = 1.0,$	200	366/373	44/5	553/614	11/4	26/4
$\beta_1 = -0.8$	500	267/274	49/3	636/721	11/2	37/0
	1000	213/221	43/6	674/771	17/1	53/1
	2000	130/143	56/1	774/853	11/1	29/2
	5000	27/41	44/2	866/954	30/2	33/1
TDL						
	100	644/658	24/3	314/337	3/0	15/2
$\alpha = 0.7, \beta_0 = -1.0,$	200	663/672	29/4	302/323	1/0	5/1
$\beta_1 = -1.2$	500	734/744	25/1	239/255	2/0	0/0
	1000	792/813	33/2	174/185	0/0	1/0
	2000	854/887	39/0	104/113	2/0	1/0
	5000	896/953	62/0	42/47	0/0	0/0
CLLGF						
	100	28/65	315/277	45/131	296/272	316/255
$\alpha = -1.1, \beta_0 = 1.0,$	200	5/16	361/353	3/15	372/365	256/251
$\beta_1 = -0.8, \theta = 0.7$	500	0/0	338/338	0/0	476/476	186/186
	1000	0/0	270/270	0/0	617/617	113/113
	2000	0/0	208/208	0/0	762/762	30/30
	5000	0/0	97/97	0/0	902/902	1/1
CLLIGF						
	100	52/79	249/207	143/302	115/97	441/315
$\alpha = -1.1, \beta_0 = 1.0,$	200	26/12	191/197	96/26	130/135	557/630
$\beta_1 = -0.8, \sigma^2 = 0.7$	500	0/0	114/113	0/1	193/193	693/693
	1000	0/0	65/65	0/0	184/184	751/751
	2000	0/0	17/17	0/0	143/143	840/840
	5000	0/0	0/0	0/0	57/57	943/943
TDLF						
	100	165/280	387/389	40/44	110/77	298/210
$\alpha = 0.4, \beta_0 = 1.5,$	200	60/186	494/420	6/7	141/123	299/264
$\beta_1 = 0.9, \theta = 0.9$	500	12/20	592/586	0/0	181/180	215/214
	1000	0/1	655/654	0/0	228/228	117/117
	2000	0/0	735/735	0/0	230/230	35/35
	5000	0/0	757/757	0/0	242/242	1/1

5. Application

5.1 Lung cancer study

To measure the annual incidence of lung cancer in Northern Ireland in one year (Wilkinson, 1995), a study was conducted between 01/10/1991 and 30/09/1992. In this period, 900 cases of lung cancer were identified. From these 900 cases a group of 20 cases were diagnosed after the patients died and in 25 patients the cause of death could not be determined. The total number of patients analyzed was 855 (95%). The observed time represents the time until death or censoring (months), 21% of the data is censored. This data set was used previously in Mackenzie (1996) and Milani (2011). We consider the metastasis covariate, which consists of three levels ("No", "Yes" and "Unknown"). Using the method described in Colosimo and Giolo (2006), the proportional hazard assumption was not observed for this covariate (Figure 5).

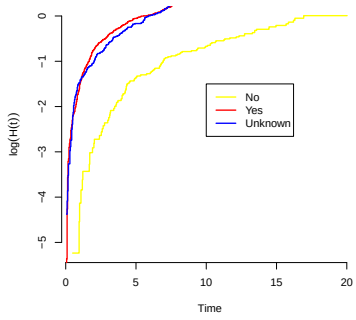

Figure 5. Proportional hazard assumption for the metastasis covariate

5.2 Models

Since the metastasis covariate has a non-proportional behaviour, we can fit CLL, CLLGF and CLLIGF models and also TDL and TDL with gamma frailty terms (TDLF). Software R was used for the optimization of the logarithm of the likelihood function and also to obtain the variance-covariance matrix. The results are shown in Tables 6 and 7.

Table 5. CLL fitted model

	CLL		
	MLE	SE	Interval
α	0.015	0.005	(0.024; 0.006)
β_0	0.543	0.044	(0.629; 0.457)
β_1	0.442	0.052	(0.341; 0.543)
β_2	-0.028	0.048	(-0.122; 0.065)

As indicated by the results presented in Tables 6 and 7, the regression coefficient associated with metastasis covariate and α, which measures the effect of time, are significant in all fitted models. The parameter that measures the variance of the frailty term is significant in all models with a frailty term. It indicates that there are covariates that were not used to fit the models, but have some influence in the survival times.

The TDL and CLL models and their extensions with frailty terms can become a survival model with cure fraction depending on the value of α. Mackenzie (1996) highlighted the

Table 6. CLLGF and CLLIGF fitted models

	CLLGF			CLLIGF		
	MLE	SE	Interval	MLE	SE	Interval
α	-0.027	0.012	(-0.051; -0.003)	-0.029	0.014	(-0.056; -0.002)
β_0	0.436	0.074	(0.291; 0.580)	0.379	0.089	(0.204; 0.555)
β_1	0.740	0.101	(0.542; 0.938)	0.779	0.121	(0.542; 1.017)
β_2	-0.069	0.088	(-0.242; 0.104)	-0.064	0.093	(-0.246; 0.118)
θ/σ^2	0.645	0.138	(0.374; 0.917)	1.130	0.379	(0.386; 1.873)

Table 7. TDL and TDLF models

	TDL			TDLF		
	MLE	SE	Interval	MLE	SE	Interval
α	-0.043	0.012	(-0.020; -0.066)	0.080	0.034	(0.014; 0.146)
β_0	-1.469	0.098	(-1.661; -1.277)	-1.302	0.148	(-1.593; -1.012)
β_1	-1.162	0.135	(-1.427; -0.897)	-1.841	0.251	(-2.332; -1.349)
β_2	0.060	0.104	(-0.144; 0.264)	0.138	0.176	(-0.207; 0.484)
θ	———	———	———	0.771	0.172	(0.433; 1.108)

fact that the survival function is improper in the TDL model when the value of α is negative. Milani (2011) observed a similar result for the TDLF model. Comparing the results presented in Tables 6 and 7, the models without a frailty term indicate a cure fraction while the models with a frailty term do not indicate this.

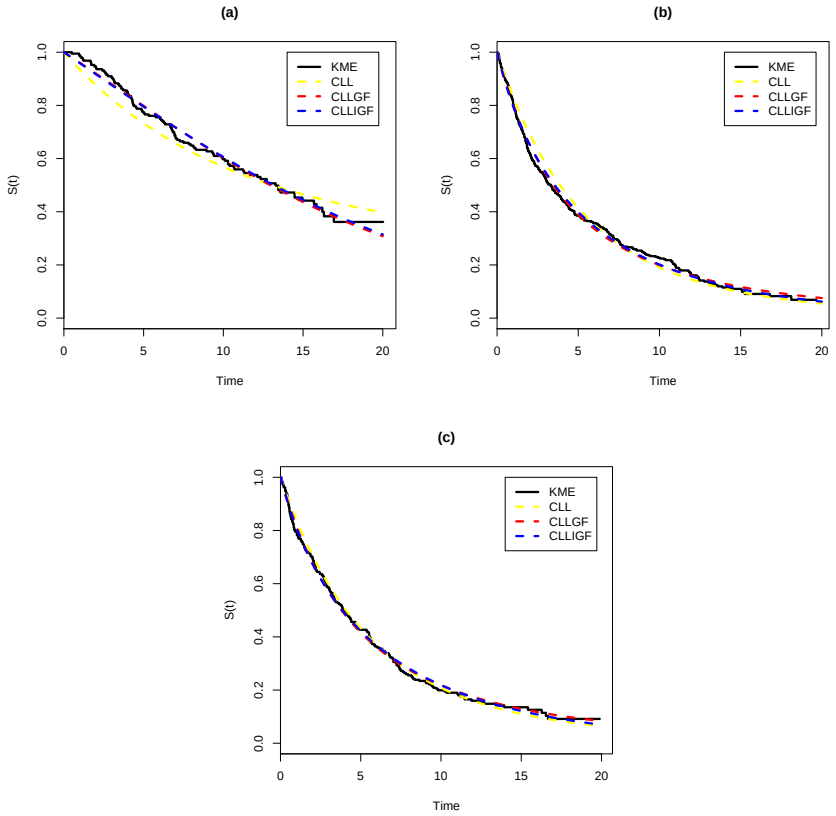

Figure 6. Curves for the survival functions for level "No" in (a), level "Yes" in (b) and level "Unknown" in (c)

The Kaplan-Meier, CLL, CLLGF and CLLIGF survival curves are presented in Figure 6. We can note from the figure that the curves associated with models with frailty are close together and for time values greater than 15 months at the level "No", the curves begin to depart from the Kaplan-Meier curve.

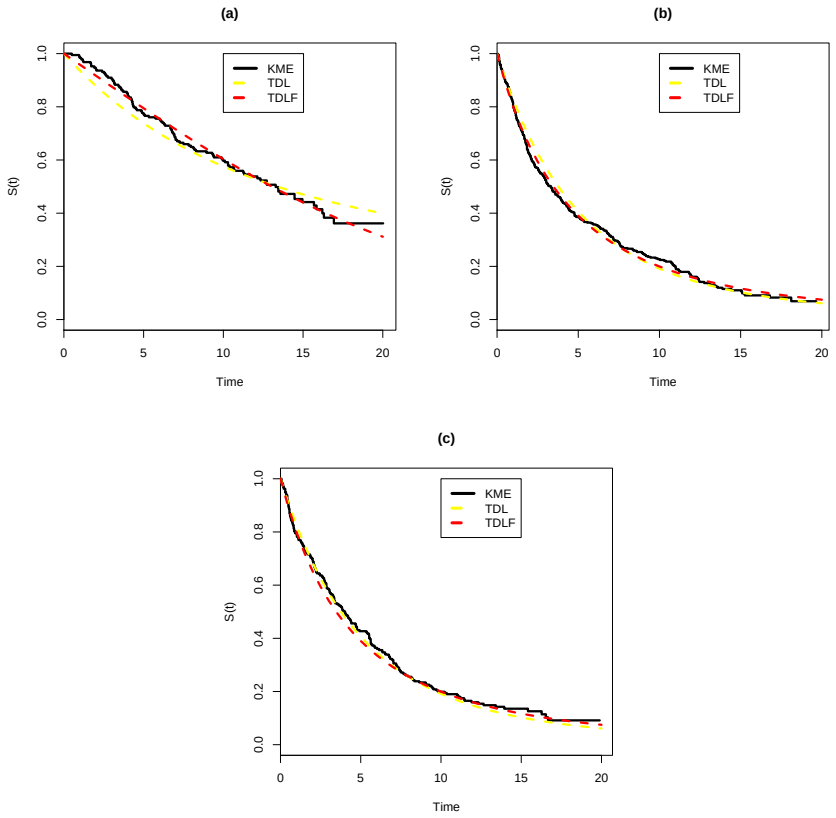

Figure 7. Curves for the survival functions for level "No" in (a), level "Yes" in (b) and level "Unknown" in (c)

The Kaplan-Meier, TDL and TDLF survival curves are presented in Figure 7. Note that TDL and TDLF curves are close to the Kaplan-Meier curve except at level "'No"' where at the end of the observed times the TDL and TDLF curves begin to depart from the Kaplan-Meier curve.

From the curves presented in Figures 6 and 7, it can be seen that the cure fraction is not evident in the levels "Yes" and "Unknown" but it is clear at level "No". The fact that the used models are or not long-term is imposed by the α value and not by the covariate values. It prevents the cure fraction estimation for some levels. Also note that the models with frailty terms are better fitted to the data than the models without frailty terms. Three model selection criteria, logarithm of the likelihood function, AIC and BIC, were used to select the model that best describes the data set (Table 8).

Table 8. Comparasion of models

model	CLL	CLLGF	CLLIGF	TDL	TDLF
log-lik	1997.002	1986.852	1986.926	1994.865	1986.915
AIC	4002.004	3983.704	3983.852	3997.730	3983.830
BIC	4020.490	4006.811	4006.959	4016.216	4006.937

From the results presented in Table 8, we can see that the CLLFG model is the best

model, according to both criteria, the AIC and the BIC. However, there is a small difference among the values of the criteria for the models with frailty term. Considering only the models without frailty term the TDL model is more indicated than the CLL model

6. Concluding Remarks

In this paper, we introduce three new models to analyze survival data, CLL (Complementary Log-Log), CLLGF (Complementary Log-Log with Gamma frailty) and CLLIGF (Complementary Log-Log with inverse gaussian frailty) finding its hazard, survival and density functions. These models can be considered alternatives to modeling data that support the assumption of non-proportional hazards. The simulation study showns that the MLEs are unbiased, the bias and SRMSE decrease when the sample size increasesand that for samples of reasonable size, the coverage probabilities are close to the nominal. In the dataset of lung cancer, in which the metastasis covariate is non-proportional, we observed the application of the new models. We note that according to the criteria AIC and BIC, the model CLLGF can be the candidate among the five fitted models as the most suited for the data. The fact that θ is significant indicates that there are factors that were not observed, but have influence on the lifetimes. We then may conclude that using the model without frailty can lead to wrong interpretations. Parameter α, which measures the effect of the time in CLL, CLLGF and CLLIGF model, is significant. Extensions of the CLL frailty model to incorporate more robust frailty distributions can be futher discussed in future research.

Acknowledgements

The authors are grateful to Gilbert MacKenzie (University of Limerick, Ireland) for providing the Northern Ireland lung cancer data. The research of Eder A. Milani was supported by CAPES-Brazil.

References

Byrd, R. H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. SIAM J. Scientific Computing 16, 1190–1208.

Ciampi, A., Etezadi-Amoli, J., 1985. A general model for testing the proportional hazards and the accelerated failure time hypotheses in the analysis of censored survival data with covariates. Communications in Statistics - Theory and Methods 14, 651–667.

Colosimo, E. A., Giolo, S. R., 2006. Anlise de Sobrevivncia Aplicada. Edgard Blcher, São Paulo.

Cox, D. R., 1972. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society B 34(2), 187–220.

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

Franco, N. M. B., 2010. Cálculo Numérico. Prentice Hall (Pearson), São Paulo.

Hougaard, P., 1995. Frailty models for survival data. Lifetime Data Analysis 1, 255–273.

Klabfleish, J. F., Prentice, R. L., 2002. The Statistical Analysis of Failure Time Data. John Wiley and Sons, New Jersey.

Louzada-Neto, F., 1997. Extended hazard regression model for reliability and survival analysis. Lifetime Data Analysis 3, 367–381.

Louzada-Neto, F., 1999. Polyhazard models for lifetime data. Biometrics 55, 1281–1285.

Mackenzie, G., 1996. Regression models for survival data: The generalized time-dependent logistic family. The Statistician 45, 21–34.

McCullagh, P., Nelder, J. A., 1898. Generalized Linear Models. Chapman & Hall/CRC, New York.

Milani, E. A., 2011. Modelo logístico generalizado dependente do tempo com fragilidade gama. Master's thesis, Universidade Federal de São Carlos - UFSCar, São Carlos.

Ruggiero, M., Lopes, V., 1997. Cálculo Numérico: Aspectos Teóricos e Computacionais. Pearson, São Paulo.

Tomazella, V. L. D., Louzada-Neto, F., da Silva, G., 2006. Bayesian modeling of recurrent events data with an additive gamma frailty distribution and a homogeneous poisson process. Journal of Statistical Theory and Applications 5, 417–429.

Wienke, A., 2011. Frailty Models in Survival Analysis. Chapman & Hall/CRC biostatistics series.

Wilkinson, P., 1995. Lung cancer in northern ireland. Master's thesis, Queen's University of Belfast, Belfast.