GOMPERTZ REGRESSION MODEL WITH GAMMA FRAILTY: A STUDY ON THE APPLICATION IN LUNG CANCER

Vera Lucia Damasceno TOMAZELLA[1]
Eder Ângelo MILANI[2]
Teresa Cristina Martins DIAS[1]

■ ABSTRACT: Survival models with frailty are used when some variables are non-available to explain the occurrence time of an event of interest. This non-availability may be considered as a random effect related to unobserved covariates, or that cannot be measured, such as environmental or genetic factors. This paper focuses on the Gamma-Gompertz (denoted by G-G) model that is one of a class of models that investigate the effects of unobservable heterogeneity. We assume that the baseline mortality rate in the G-G model is the Gompertz model, in which mortality increases exponentially with age and the frailty is a fixed property of the individual, and the distribution of frailty is a gamma distribution. The proposed methodology uses the Laplace transform to find the unconditional survival function in the individual frailty. Estimation is based on maximum likelihood methods and this distribution is compared with its particular case. A simulation study examines the bias, the mean squared errors and the coverage probabilities considering various samples sizes and censored data. A real example with lung cancer data illustrates the applicability of the methodology, where we compared the G-G and without frailty models via criteria which select the best fitted model to the data.

■ KEYWORDS: Gamma-Gompertz model; proportional hazard; survival model.

1 Introduction

Vaupel et al. (1979) introduced the term frailty to indicate that different individuals are at risks even though on the surface they may appear to be quite

[1]Universidade Federal de São Carlos - UFSCAR, Departamento de Estatística, CEP: 13.565-905, São Carlos, SP, Brasil. E-mail: *vera@ufscar.br*; *tcristina.md@gmail.com*
[2]Universidade Federal de Goiás - UFG, Instituto de Matemática e Estatística, CEP: 74.690-900 Goiânia, GO, Brazil. E-mail: *edinhomilani@hotmail.com*

similar with respect to measurable attributes such as age, gender, weight, etc. They used the term frailty to represent an unobservable random effect shared by subjects with similar (unmeasured) risks in the analysis of mortality rates. A random effect describes excess risk or frailty for distinct categories, such as individuals or families, over and above any measured covariates. Thus random effects, or frailty models, have been introduced into the statistical literature in an attempt to account for the existence of unmeasured attributes such as genotype that do introduce heterogeneity into a study population. A common approach to the analysis of survival data is to assume a homogeneous population of individuals with the same covariate structure. However, it is clear that individuals identical in many respects such as age, sex and treatment may differ in unmeasured ways, only because of genotypical differences. It is easy to see that it is important to consider the effect of ignoring frailty in any study where the existence of such heterogeneity may be present.

More formally, a heterogeneous population can be sometimes modeled as a mixture problem with an underlying random variable called frailty. This random effect or frailty is introduced in the baseline hazard rate (HR) additively or multiplicatively. Several authors have studied the use of multiplicative frailty models, which represent a generalization of the Cox model (COX, 1972). Andersen (1993) and Houggard (1995) presented a review of the multiplicative frailty models in the classical perspective, whereas Sinha and Dey (1997) presented a review of these models under the Bayesian point of view. Some authors have studied models with univariate frailty. For example, Aalen and Tretli (1999) applied the compound-Poisson distribution to data from testicular cancer, Henderson and Oman (1999) studied the consequence of ignoring the frailty in the fitting, Tomazella et al. (2008) presented an approach involving objective Bayesian reference analysis to the frailty model with survival time univariate, Hanagal and Sharma (2012) considered the shared gamma frailty model with Gompertz distribution as baseline hazard for bivariate survival times, Sharma and Hanagal (2014) proposed frailty regression models in Gompertz mixture distributions and assume the distribution of frailty as gamma or inverse Gaussian or positive stable or power variance function distribution and, Milani et al. (2015) proposed a frailty model with non-proportional hazard.

Suppose that T is the occurence time of a event of the subject (for example, the time to infection for a kidney patient using portable dialysis) and the x is a covariate; then, the probability density of T might be modeled conditional on v, an unobserved non-measurable random variable, called frailty, which is intended to allow for individual variation. This representation can be symbolized by $f(t; x, v)$. Under this representation, the occurence-time distribution can be considered to be continuous mixture induced by the frailty v.

The frailty term of the model is random and a distribution must be assumed for it. Due to the way as the frailty term acts on the HR, natural candidates to the frailty distribution must be supposed as continuous and time independent, such as gamma, inverse Gaussian, log-normal and Weibull distributions (HOUGGARD, 1995). In this paper, we focus on gamma distribution and use the G-G frailty model. The frailty in the frailty model is assumed to follow a gamma distribution

and baseline HR is the Gompertz model.

The Gompertz distribution is widely used to fit data from clinical trials and construct life tables in actuarial area. Various authors have studied the parameters estimation of this model. Among others, Garg et al. (1970) obtained the maximum likelihood estimators and properties for the parameters and Lenart (2012) presented a revision about Gompertz distribution. The G-G model is one of a class of models that investigate the effect of unobserved heterogeneity that is either unobservable or unobserved on mortality rates.

Our goal is analyze a dataset that represents a patient's lifetime with lung cancer when we adopt the G-G regression model in a possible cure rate scenario. In this methodology, we use the Laplace transform of the frailty density to obtain the population (or unconditional) survival function. The use of this transformation makes it easier to obtain survival function.

This paper is organized as follows. In Section 2 we present review of frailty model, in Section 3 we present Gompertz regression model and in Section 4 we present Gompertz regression model with gamma frailty and the estimation about the parameters of this model. In section 5 we present a simulation study of the proposed model considering right-censored and no censored data to some samples sizes and in Section 6 we present an application on a real lung cancer dataset. Finally, in Section 7 we make some concluding remarks.

2 Background

2.1 Frailty models

Consider an unobserved source of heterogeneity, which is not readily captured by a covariate on a univariate frailty model. It extends the Cox model, such that the hazard function (HF) of a patient depends on an unobservable value of the random variable V, which acts multiplicatively on the baseline HF. Therefore, the conditional HR of V random variable to pacient i available at time t, given $V = v_i$, is given by

$$h_{T|V=v_i}(t) = v_i h_0(t), \quad i = 1, \ldots, n, \quad t > 0, \tag{1}$$

where v_i is the frailty of the patient i and $h_0(t)$ is a baseline HF. Note that (1) is known as the Clayton model (CLAYTON, 1991). From (1), note that the HR of the patient i decreases if $v_i < 1$ and increases if $v_i > 1$. The corresponding conditional survival function (SF) can be obtained from (1) as

$$S_{T|V=v_i}(t) = \exp(-v_i H_0(t)), \quad i = 1, \ldots, n, \quad t > 0, \tag{2}$$

where $H_0(t) = \int_0^t h_0(s)\mathrm{d}s$ is the baseline cumulative hazard function.

Let $(\boldsymbol{t}, \boldsymbol{\delta})$ be the observed data for a sample of size n, where t_i is the occurrence-time of the event of interest and δ_i is the indicator of censoring, that is, $\delta_i = 1$ if

the observation is the time to the event of interest and $\delta_i = 0$ if it is right censored, for $i = 1, \ldots, n$. Then, from (1) and (2), the corresponding likelihood function is

$$L(\boldsymbol{\mu}; \boldsymbol{t}, \boldsymbol{v}, \boldsymbol{\delta}) = \prod_{i=1}^{n} \left(v_i h_0(t_i)\right)^{\delta_i} \exp\left(-v_i H_0(t_i)\right), \tag{3}$$

where $\boldsymbol{\mu}$ is the vector of parameters, $\boldsymbol{t} = (t_1, \ldots, t_n)$, $\boldsymbol{v} = (v_1, \ldots, v_n)$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$. Now, conditional on the unobserved frailties $\boldsymbol{v}$, the likelihood function given in (3) forms the basis for the parameters estimation. The frailties must be integrated out (in closed form or by numerical or stochastic integration, depending on the frailty distribution) to get a likelihood function not depending on unobserved quantities of the type

$$L(\boldsymbol{\mu}; \boldsymbol{t}, \boldsymbol{\delta}) = \prod_{i=1}^{n} \left(h_T(t_i)\right)^{\delta_i} S_T(t_i). \tag{4}$$

2.2 Unconditional hazard and survival functions

The unconditional (population) SF of T can be obtained by integrating $S_{T|V=v_i}(t)$ given in (2) on the frailty v. This function may be viewed as the 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 can be obtained with the Laplace transform (HOUGAARD, 1984). Then, when seeking distributions for the frailty variable V, it is natural to use frailty distributions with an explicit Laplace transform, because it facilitates the use of traditional maximum likelihood method for parameter estimation. To obtain the unconditional SF, we need to integrate out the frailty component as

$$S_T(t) = \int_0^\infty S_{T|V}(t) f_V(v) \mathrm{d}v, \tag{5}$$

where $f_V(v)$ is the probability density function (PDF) of the V and $S_{T|V}(t)$ is the conditional SF given in (2).

In general, the Laplace transform of real argument s of a function $f(x)$ is given by

$$Q(s) = \int_0^\infty \exp(-sx) f(x) \mathrm{d}x. \tag{6}$$

Let $f(\cdot) = f_v(\cdot)$ be the frailty PDF and $s = H_0(t)$. Then, according to (6), we obtained the Laplace transform of the unconditional SF as

$$S_T(t) = \int_0^\infty \exp(-vH_0(t)) f_V(v) \mathrm{d}v = Q(H_0(t)). \tag{7}$$

Note that (7) conducts to the same form as the unconditional SF given in (5) (see VAUPEL et al., 1979 and WIENKE, 2011). The frailties random variables v_i

are usually assumed to be independent and identically distributed. As mentioned, the frailty distribution can be gamma, inverse Gaussian or Weibull, which have simple Laplace transforms and then are convenient to use. In this paper we considered a reparametrized version of the gamma distribution traditionally used in frailty models.

3 Gompertz regression model

The HF of the Gompertz model, in which mortality increases exponentially with age t, is given by

$$h(t|b,\alpha) = be^{\alpha t}, \tag{8}$$

where b denotes the level of the force of mortality at age $t = 0$ and α the rate of aging.

We considered a continuous random variable T with a Gompertz PDF with location parameter b and shape parameter α,

$$f(t|b,\alpha) = b\exp\left(\alpha t - \frac{b}{\alpha}\left(e^{\alpha t}-1\right)\right), \;\; t>0,\; \alpha>0 \text{ e } b>0.$$

The truncated distribution yields a proper density function by rescaling the α parameter to correspond to $t = 0$ (GARG et al., 1970 and LENART, 2012). The distribution function is

$$F(t|b,\alpha) = 1-\exp\left(-\frac{b}{\alpha}\left(e^{\alpha t}-1\right)\right).$$

Considering $b = \exp(\boldsymbol{x}'\boldsymbol{\beta})$ in (8) we have the Gompertz regression model or time-dependent proportional hazard model. This model is defined by the HF given by

$$h(t|\alpha,\boldsymbol{\beta}) = \exp(\alpha t + \boldsymbol{x}'\boldsymbol{\beta}), \tag{9}$$

where α is a measure of the time effect, $\boldsymbol{\beta}'=(\beta_0,\beta_1,\ldots,\beta_k)$ is a vector of $k+1$ unknown parameters measuring the influence of the k covariates $\boldsymbol{x}' = (1,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k)$ and, t represents the univariate survival time of a unit or individual.

The behavior of the HF (9) takes several forms, according to the value of α: for $\alpha > 0$, the hazard function is increasing; for $\alpha < 0$, the hazard function is decreasing and for $\alpha = 0$, the hazard function is constant. The Figure 1 (left) shows some examples of possible shapes of the hazard function.

When the survival times of the n individuals are observed, the ratio of the hazard function of two individuals, i and j, with $i \neq j$ and $i,j = 1,\ldots,n$, with different covariates vector is given by

$$\frac{h_i(t|\boldsymbol{x}_i)}{h_j(t|\boldsymbol{x}_j)} = \frac{\exp(\alpha t + \boldsymbol{x}_i'\boldsymbol{\beta})}{\exp(\alpha t + \boldsymbol{x}_j'\boldsymbol{\beta})} = \exp[(\boldsymbol{x}_i - \boldsymbol{x}_j)'\boldsymbol{\beta}]. \tag{10}$$

Note that the time effect disappears in equation (10) and hence the proportionality becomes evident.

From equation (9) the SF is given by

$$S\left(t|\alpha,\boldsymbol{\beta}\right)=\exp\left\{\frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{\alpha}\left[1-e^{\alpha t}\right]\right\}. \tag{11}$$

Observe that the function in (11) also has its behavior determined by the value of α. For $\alpha>0$, $S(0|\alpha,\boldsymbol{\beta})=1$ and $S(\infty|\alpha,\boldsymbol{\beta})=\lim\limits_{t\longrightarrow\infty}S(t|\alpha,\boldsymbol{\beta})=0$, in other words, the survival function is proper. For $\alpha<0$, $S(0|\alpha,\boldsymbol{\beta})=1$ and $S(\infty|\alpha,\boldsymbol{\beta})\neq 0$, the survival function is improper, that is, when $\alpha<0$ we have a model for cure rate or long duration, with the cure rate, p, given by

$$p=\exp\left\{\frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{\alpha}\right\}.$$

Some examples of hazard functions (left) and survival functions are shown in Figure 1 (right).

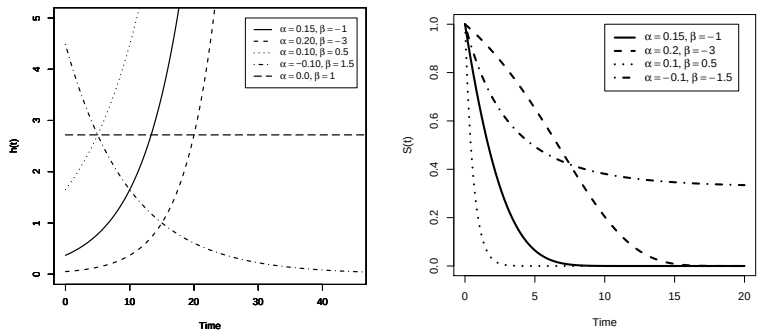

Figure 1 - Different forms of the hazard function (left) and survival function (right) to the Gompertz regression model.

Using the equations (9) and (11), the PDF of T is given by

$$f\left(t|\alpha,\boldsymbol{\beta}\right)=\exp\left\{\alpha t+\boldsymbol{x}'\boldsymbol{\beta}+\frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{\alpha}\left(1-e^{\alpha t}\right)\right\},$$

when the survival function is proper.

Let $(t_i,\boldsymbol{x}_i,\delta_i)$ be n the observed times, δ_i is an indicator variable and $\boldsymbol{x}_i'$ is a covariates vector, $i=1,\ldots,n$. The likelihood function for right-censored data is

given by

$$L\left(\alpha,\boldsymbol{\beta}|\boldsymbol{t},\boldsymbol{x},\boldsymbol{\delta}\right)=\prod_{i=1}^{n}\left[\exp\left(\alpha t_i+\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right]^{\delta_i}\exp\left\{\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}\left(1-e^{\alpha t_i}\right)\right\}. \tag{12}$$

The maximum likelihood estimators (MLEs) are obtained by direct maximization of equation (12) or through the log-likelihood function. The asymptotic confidence intervals are obtained assuming asymptotic normality of the MLEs.

4 Gompertz regression model with gamma frailty

Considering the frailty model given in (1) and the equation (9), the HF of the ith individual with the multiplicative frailty term $v_i (v_i > 0)$ is given by,

$$h_i(t|\alpha,\boldsymbol{\beta},v_i)=v_i\exp(\alpha t+\boldsymbol{x}_i'\boldsymbol{\beta}), \tag{13}$$

interpreted as the conditional hazard function of the ith individual given v_i and the respective conditional SF is given by

$$S_i(t|\alpha,\boldsymbol{\beta},v_i)=\exp\left\{-\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}(e^{\alpha t}-1)v_i}{\alpha}\right\},\quad i=1,\ldots,n. \tag{14}$$

In models with multiplicative frailty, we are considering that different individuals have different frailties. Then, the individuals who present the highest values of variable v_i tend to die earlier than the individuals who present the lowest values of the same variable.

The frailty model not only explains the heterogeneity among individuals, but it also allows to assess the covariates effect that for some reason were not considered at the fitting.

The value v is not observed, which is why we assume that v is an observation of the random variable V with a given probability density function. In the literature the gamma, lognormal, Weibull and inverse Gaussian distributions are the most used (HOUGAARD, 1995). In this paper we considered that V has a gamma distribution with parameters $\tau > 0$ and $\eta > 0$, $\mathrm{G}(\tau,\eta)$, with density function written as

$$f_V(v;\tau,\eta)=\frac{\eta^{\tau}}{\Gamma(\tau)}v^{(\tau-1)}\exp\left(-v\eta\right). \tag{15}$$

Considering univariate times, if we built the likelihood function using the hazard and survival functions given in (13) and (14), respectively, we would have more parameters than observations, so we need to calculate the hazard and unconditional survival functions. In the context of proportional hazard, according to Elbers and Ridder (1982), when working with frailty it is necessary that the

random effect distribution has finite mean for the model to be identifiable. This way, in order to keep the identifiability of the model it is convenient to take the distribution with mean 1.

Thus, we assume the gamma distribution given in (15) with parameter τ=$1/\theta$ and η=$1/\theta$, where E(V)=1 and Var(V)=θ.

To get the unconditional SF we need to calculate

$$S(t) = \int_0^\infty S(t|\alpha, \boldsymbol{\beta}, v) f(v) dv,$$

where $f(v)$ is the PDF in (15). In order to calculate the unconditional SF we use the Laplace transform since both have the same shape. The Laplace transform of the gamma distribution (15) with parameters $(1/\theta, 1/\theta)$ (WIENKE, 2011) and considering s a real argument, is given by

$$Q(s) = (1 + \theta s)^{-1/\theta}. \tag{16}$$

Substituting $s = H(t)$ in the equation (16), we obtain the unconditional SF, given by

$$S(t|\alpha, \boldsymbol{\beta}, \theta) = \left[1 + \frac{\theta}{\alpha} e^{\boldsymbol{x}'\beta} \left(e^{\alpha t} - 1\right)\right]^{-1/\theta}. \tag{17}$$

The behavior of the SF is determined by the value α. For $\alpha > 0$ the survival function is proper and for $\alpha < 0$ it is improper, then we have a long duration model with the cure rate, p, given by

$$p = \left(1 - \frac{\theta}{\alpha} e^{\boldsymbol{x}'\beta}\right)^{-1/\theta}.$$

From equation (17) we get the correspondent hazard function, given by

$$h(t|\alpha, \boldsymbol{\beta}, \theta) = \frac{e^{\alpha t + \boldsymbol{x}'\beta}}{1 + \frac{\theta}{\alpha} e^{\boldsymbol{x}'\beta} \left(e^{\alpha t} - 1\right)}. \tag{18}$$

In this case we have hazard function of the G-G model that is an attractive explanation for the widely observed pattern of decelerating increase in mortality with age, in both humans and other species (e.g., VAUPEL et al., 1979).

Some examples of hazard (18) and survival (17) functions are shown in Figure 2.

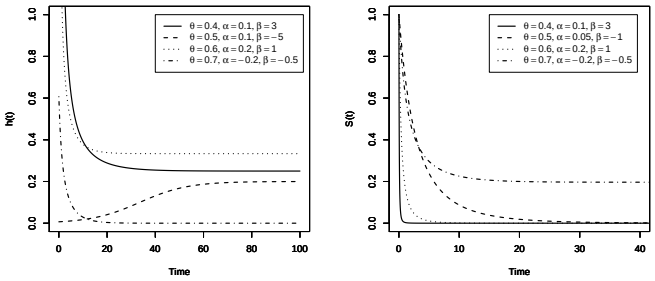

Figure 2 - Different forms of the hazard function (left) and survival function (right) for the G-G model.

4.1 Inference

Let $(t_i, \boldsymbol{x}_i, \delta_i)$ be n the observed times, δ_i is an indicator variable and $\boldsymbol{x}_i'$ is a covariates vector, $i = 1, \ldots, n$. The likelihood function for right-censored data, constructed from the equations (17) and (18), is given by

$$\begin{aligned} L(\alpha, \boldsymbol{\beta}, \theta | \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\delta}) &= \prod_{i=1}^{n} [h(t_i; \alpha, \boldsymbol{\beta}, \theta)]^{\delta_i} [S(t_i; \alpha, \boldsymbol{\beta}, \theta)] \\ &= \prod_{i=1}^{n} \left[\frac{e^{\alpha t_i + \boldsymbol{x}_i'\boldsymbol{\beta}}}{1 + \frac{\theta}{\alpha} e^{\boldsymbol{x}_i'\boldsymbol{\beta}} \left(e^{\alpha t_i} - 1\right)} \right]^{\delta_i} \left[1 + \frac{\theta}{\alpha} \left[e^{\boldsymbol{x}_i'\boldsymbol{\beta}} \left(e^{\alpha t_i} - 1\right) \right] \right]^{-1/\theta}, \end{aligned} \tag{19}$$

where $\boldsymbol{t} = (t_1, \ldots, t_n)$, $\boldsymbol{x}' = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ is arandom variable censoring indicator. The log-likelihood function (from (19)) is given by

$$\begin{aligned} \log(L(\alpha, \boldsymbol{\beta}, \theta | \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\delta})) = &- \sum_{i=1}^{n} \frac{\log\left(\frac{\theta\left(e^{\alpha t_i} - 1\right) e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha} + 1 \right)}{\theta} \\ &- \sum_{i=1}^{n} \delta_i \log\left(\frac{\theta \left(e^{\alpha t_i} - 1\right) e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha} + 1 \right) + \sum_{i=1}^{n} \delta_i \log\left(e^{\alpha t_i + \boldsymbol{x}_i'\boldsymbol{\beta}} \right). \end{aligned} \tag{20}$$

The MLEs are obtained by direct maximization of equation (19) or by maximization of (20). In the case of uncensored times, the log of the likelihood function and the first and second derivatives of the α, β and θ parameters are shown in Appendix.

The asymptotic confidence intervals are obtained assuming asymptotic normality of the maximum likelihood estimators. The comparison between the

Gompertz regression model and Gompertz regression model with gamma frailty is made by considering the Akaike information criterion (AIC) (AKAIKE, 1974) and the Bayesian information criterion (BIC) (SCHWARZ, 1978). The lowest AIC and BIC indicate the best fitted model to data.

5 Simulation study

In this work, the main concern of the simulation study is to assess the mean absolute bias (MAB) and mean squared error (MSE) of the MLEs, as well as the coverage probabilities of the asymptotic confidence intervals for the parameters of the frailty model. We generated 5,000 samples for each sample size ($n = 50, 100, 200$ and 300). The parameters were fixed at $\alpha = 0.1$, $\beta_0 = 1$, $\beta_1 = -1.8$ and $\theta = 0.5$ and the dummy covariate was generated from a Bernoulli distribution with success probability equal to 0.6. The choice of the parameters was made considering the form of the risk function and for the $\alpha > 0$ the survival function is proper. The right-censored times were generated from an exponential distribution with mean equals to 16.94 for 10% of censoring and 3.70 for 30% of censoring. The software R (R Core Team, 2017) was used in the analysis.

For each sample we obtained the maximum likelihhod estimatives and the asymptotic 95% confidence intervals. Using these values, we calculated MAB, MSE and the coverage probability (CP). For example, for the parameter α there is

$$\text{MAB}(\alpha) = \frac{\sum_{j=1}^{5000} |\alpha^* - \widehat{\alpha}^j|}{5000},$$

where α^* is the true value of the parameter α and $\widehat{\alpha}^j$ is the maximum likelihood estimate of α in the jth sample,

$$\text{MSE}(\alpha) = \frac{\sum_{j=1}^{5000} (\alpha^* - \widehat{\alpha}^j)^2}{4999},$$

and CP is given by the quotient between the number of intervals containing the true parameter value and the total number of intervals constructed. The MAB and MSE are shown in the Table 1 and the CP in Figure 3.

Table 1 - MAB and MSE of the MLEs for 0% / 10% / 30% of censoring

Size	Censoring	MAB				MSE			
		α	β_0	β_1	θ	α	β_0	β_1	θ
n=50	0	0.1669	0.3336	0.3913	0.4250	0.0585	0.1791	0.2508	0.3159
	10	0.1973	0.3416	0.4186	0.4477	0.0743	0.1889	0.2829	0.3372
	30	0.2863	0.3485	0.4645	0.4958	0.1428	0.1940	0.3443	0.3947
n=100	0	0.1126	0.2369	0.2744	0.2892	0.0264	0.0900	0.1221	0.1494
	10	0.1346	0.2440	0.2948	0.3116	0.0365	0.0956	0.1434	0.1700
	30	0.2108	0.2572	0.3369	0.3770	0.0833	0.1062	0.1843	0.2495
n=200	0	0.0736	0.1597	0.1916	0.1887	0.0097	0.0402	0.0577	0.0583
	10	0.0877	0.1685	0.2036	0.2022	0.0139	0.0443	0.0661	0.0685
	30	0.1389	0.1759	0.2299	0.2473	0.0343	0.0487	0.0847	0.1036
n=300	0	0.0576	0.1291	0.1518	0.1484	0.0056	0.0265	0.0370	0.0357
	10	0.0703	0.1315	0.1631	0.1607	0.0084	0.0274	0.0421	0.0420
	30	0.1101	0.1387	0.1850	0.1956	0.0207	0.0306	0.0539	0.0631

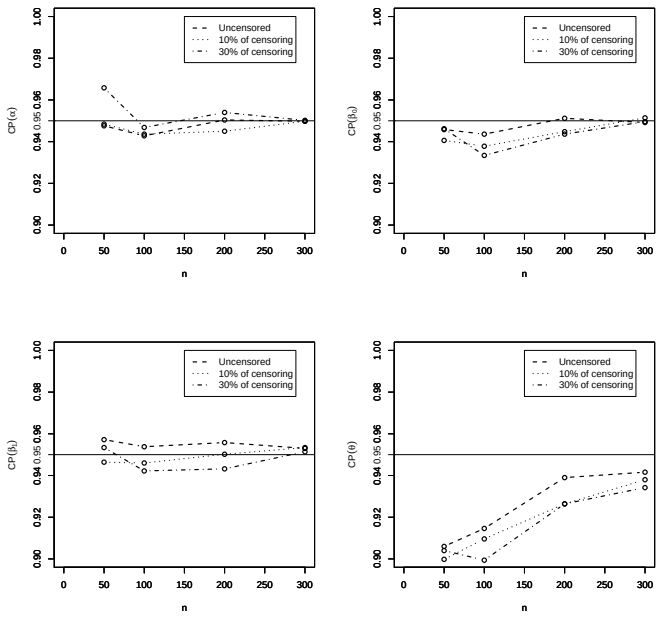

Figure 3 - Coverage probability of the asymptotic 95% confidence intervals for α, β_0, β_1 and θ.

We observed that both MAB and MSE metrics decrease with increasing sample size, for the three scenarios of censoring. We also noted that when the amount of censoring increases, the value of the metrics also increases. Comparing the values of the metrics of the parameters β_0 and β_1 with the values of α and θ, we observed that, the relative increase of the parameters that measure the effect of covariates is

almost always smaller than the other parameters, when increasing the amount of censoring present in the sample.

Regardless of the amount of censoring in the sample, the coverage probabilities of parameters seem to converge to the nominal level. We observed that the convergence of the coverage probability of the parameter θ seems to be slower than the coverage probability of other parameters.

6 Lung cancer data

To illustrate the applicability of the proposed model, we adopted the dataset of the annual incidence of lung cancer in Northern Ireland, between 01/01/1991 to 09/30/1992 (WILKINSON, 1995). In this period, 900 cases of lung cancer were identified, however, we excluded all individuals with missing information on some covariates from the analysis, resulting in 751 patients to have their lifetime analyzed (in months).

We considered for the fitting only the categorical covariates the sodium level (X_1) with the categories $<$ 136mmol/l and $>=$136mmol/l and, albumen level (X_2), with categories $<$35g/l and $>=$35g/l. We verified the assumption of proportional hazards of these covariates using the graphical method presented in Colosimo and Giolo (2006). The result of the method is shown in Figure 4 and it is possible to note that both covariates show proportional hazard.

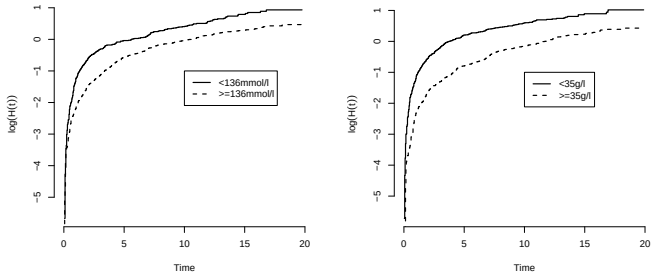

Figure 4 - Proportional hazard assumption for the sodium level (left) and albumen level (right) covariates.

We fitted the Gompertz regression and G-G models for the dataset, using the possible combinations of the covariates (X_1 only, X_2 only, and, X_1 with X_2, Tables 4, 5 and 3). We adopted the criteria AIC and BIC to select which model best fitted the data. The results are shown in Table 2. We observed that the model that best fits the data in both criteria is the Gompertz regression model with gamma frailty with the covariates X_1 and X_2 because the values of these criteria are the lowest in these models. Comparing only the scenarios with the same covariates adopted, the model with frailty (G-G) is preferred in all the settings.

Table 2 - Results of the criteria for the fitted models

Model	AIC	BIC
Gompertz with X_1 only	3,570.68	3,584.55
G-G with X_1 only	3,564.01	3,582.49
Gompertz with X_2 only	3,519.90	3,533.76
G-G with X_2 only	3,504.68	3,523.17
Gompertz with X_1 and X_2	3,503.77	3,522.25
G-G with X_1 and X_2	**3,486.51**	**3,509.61**

The maximum likelihood estimates and asymptotic 95% confidence intervals of the Gompertz and G-G models, with covariates X_1 and X_2 are presented in Table 3. We observed changes in the maximum likelihood estimates of the parameters that measure the effect of time and of the covariates, for the models with and without frailty. We also noted, that all parameters are significant, including the parameter θ, which measures the heterogeneity of the individuals.

Table 3 - The results of the fit of the without and with frailties models

	Model with frailty		Model without frailty	
Parameter	MLE	CI	MLE	CI
α	0.0674	(0.0046; 0.1303)	-0.0436	(-0.0648; -0.0224)
β_0	-0.7826	(-1.0630; -0.5022)	-1.2114	(-1.3722; -1.0506)
β_1	-0.5871	(-0.8581; -0.3160)	-0.3718	(-0.5413; -0.2022)
β_2	-1.1732	(-1.4821; -0.8643)	-0.7277	(-0.8969; -0.5585)
θ	0.8946	(0.3972; 1.3919)		

Table 4 - The results of the fit of the without and with frailties models with X_1 only

	Model with frailty		Model without frailty	
Parameter	MLE	CI	MLE	CI
α	0.0680	(-0.0286; 0.1645)	-0.0535	(-0.0748; -0.0322)
β_0	-1.2032	(-1.4779; -0.9284)	-1.5157	(-1.6675; -1.3639)
β_1	-0.9001	(-1.2286; -0.5716)	-0.5397	(-0.7042; -0.3752)
θ	1.0843	(0.2443; 1.9243)		

Table 5 - The results of the fit of the without and with frailties models with X_2 only

	Model with frailty		Model without frailty	
Parameter	MLE	CI	MLE	CI
α	0.0643	(-0.0007; 0.1292)	-0.0461	(-0.0673; -0.0250)
β_0	-1.0481	(-1.2852; -0.8111)	-1.3791	(-1.5241; -1.2341)
β_2	-1.3093	(-1.6319; -0.9867)	-0.8158	(-0.9799; -0.6517)
θ	0.9038	(0.3804; 1.4271)		

We observed that in the fitted frailty model, there is a reduction of heterogeneity present in the data. For example, in the various scenarios (Table 2) when we fitted the frailty model with X_1 only covariate (Table 4), we got $\theta = 1.08$, when we fitted the model without frailty with X_2 only covariate, we got $\theta = 0.90$ (Table 5), and when we used the two covariates with X_1 and X_2 in the frailty model, we got $\theta = 0.89$ (Table 3).

In Figure 5 we presented the survival functions estimated by Kaplan-Meier (KME) (KAPLAN and MEIER, 1958) and the frailty model. We observed through the graphics that both curves are close, indicating that the G-G model showed a good fit for the data.

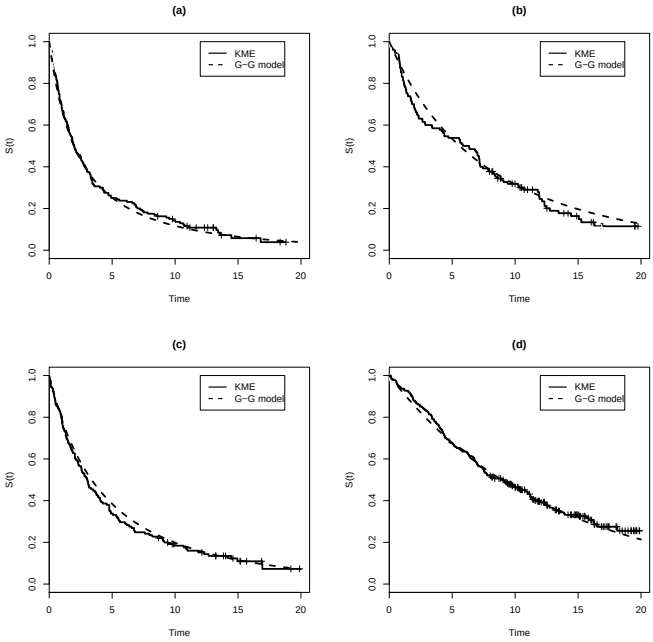

Figure 5 - The survival functions estimated by Kaplan-Meier and G-G model in (a) sodium level <136mmol/l and albumen level <35g/l, (b) sodium level >=136mmol/l and albumen level <35g/l, (c) sodium level >=136mmol/l and albumen level <35g/l and (d) sodium level >=136mmol/l and albumen level >=35g/l.

7 Final comments

In this paper we have studied a model where the gamma distribution is employed, in the Gompertz regression model, to describe the unobserved heterogeneity. We have explicitly derived the unconditional hazard rate and the survival functions using the Laplace transformation. To study this model, we presented a simulation study and a real example on lung cancer, which is compared to the modeling without frailty via selection criteria to determine which model best fits the data. More specifically, in the simulation study we considered the presence of frailties, as well as different percentage of censured data (0%, 10% and 30%) and samples sizes ($n = 50, 100, 200$ and 300). The metrics used to compare the adjusted values with real values are MAB and MSE, and we observed that when the censoring percentage is fixed and the sample size is increased, measures decrease. In addition, when n increased we observed that the estimates for parameters were very close to the real values. This fact occurred in all the studied. We noted that for the parameters α, β_1 and β_0, the CP is very close to the nominal for n greater than or

equal to 200. However, for the parameter θ, the CP has only achieved this for $n = 300$, in all studied censoring levels. In the model with fragility, the simulation study showed good properties of MLEs, giving which grants us confidence in stating that the estimation of the effect of time and covariates is important and make it possible to explain the data more accurately.

In case a lung cancer data, the use of the Gompertz regression model with gamma fragility, explained the heterogeneity present in the data when there are risks factors not observed. Using the criteria AIC and BIC, there is evidence in favour of this model, when compared to model without frailty (Figure 5). That is, the G-G model with sodium level (X_1) and albumen level (X_2) covariates fits better the dataset. Also, the use of important covariates on modeling causes decreases in heterogeneity, which is showed by parameter θ. It is important to note that in both cases, simulation study and application, the G-G model best describes the behavior of data and captures the fragility. Also, we focus on the $\alpha > 0$ case, that is, the survival function is proper. Future work may be performed for the case where the survival function is improper, that is, in long-term models.

Acknowledgments

We thank the Editors and two Referees for their comments and suggestions.

DIAS, T. C. M. Modelo de regressão de Gompertz com fragilidade Gama: Um estudo de caso de câncer de pulmão. *Rev. Bras. Biom.,* Lavras, v.36, n.4, p.860-879, 2018.

- *RESUMO: Modelos de sobrevivência com fragilidade são usados quando alguma variável não está disponível para explicar o tempo de ocorrência de um evento de interesse. Esta não disponibilidade pode ser considerada como um efeito aleatório relacionado a covariáveis não observadas, ou que não podem ser medidas, como fatores ambientais ou genéticos. Este artigo enfoca o modelo Gama-Gompertz (denotado por G-G), que pertence a uma classe de modelos que investigam os efeitos da heterogeneidade não observável. Assumimos que a taxa de mortalidade basal no modelo G-G é o modelo de Gompertz, em que a mortalidade aumenta exponencialmente com a idade e a fragilidade é uma característica fixa do indivíduo e a distribuição da fragilidade é uma distribuição gama. A metodologia proposta utiliza a transformada de Laplace para encontrar a função de sobrevivência incondicional na fragilidade individual. A estimativa é baseada em métodos de máxima verossimilhança e esta distribuição é comparada com o seu caso particular. Um estudo de simulação examina o viés, erros quadráticos médios e probabilidades de cobertura considerando vários tamanhos de amostras e dados censurados. Um exemplo real com dados sobre câncer de pulmão ilustra a aplicabilidade da metodologia, em que comparamos os modelos G-G e sem fragilidade através de critérios que selecionam o modelo mais adequado aos dados.*

- *PALAVRAS-CHAVE: Modelo Gama-Gompertz; risco proporcional; modelo de sobrevivência.*

References

AALEN, O. O.; BORGAN, O.; GJESSING, H. K. *Survival and event history analysis*. New York: Springer, 2008.

AALEN, O. O.; TRETLI, S. Analyzing incidence of testis cancer by means of a frailty model. *Cancer Causes and Control*, v.10, p.285-292, 1999.

AKAIKE, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, v.19, p.716-723, 1974.

ANDERSEN, P. *Statistical models based on counting processes*. New York: Springer, 1993.

CALSAVARA, V. F.; TOMAZELLA, V. L. D.; FOGO, J. C. The effect of frailty term in the standard mixture model. *Chilean Journal of Statistics* v.4, p.95-109, 2013.

CLAYTON, D. G. A Monte Carlo Method for Bayesian Inference in Frailty Models. *Biometrics*, v.47, p.467-485, 1991.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. São Paulo: Edgard Blucher, 2006.

COX, D. R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, v.34, n.2, p.187-220, 1972.

ELBERS, C; RIDDER, G. True and spurious duration dependence: The identifiability of the proportional hazard model. *The Review of Economic Studies*, v.49, p.403-409, 1982.

GARG, M.; RAO, B.; REDMOND, C. Maximum-likelihood estimation of the parameters of the Gompertz survival function. *Journal of the Royal Statistical Society - Series C*, v.19, n.2, p.152-159, 1970.

HANAGAL, D. D.; SHARMA, R. Bayesian estimation of parameters for the bivariate Gompertz regression model with shared gamma frailty under random censoring. *Statistics and Probability Letters*, v.82, p.1310-1317, 2012.

HENDERSON, R.; OMAN, P. Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society B*, v.61, p.367-379, 1999.

HOUGAARD, P. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, v.71, p.75-83, 1984.

HOUGAARD, P. Frailty models for survival data. *Lifetime Data Analysis*, v.1, p.255-273, 1995.

KAPLAN, E. L., MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v.53, p.457-481, 1958.

KLEIN, J. P.; MOESCHBERGER, M. L. *Survival Analysis: techniques for Censored and Truncated Data*. New York: Springer, 2003.

LENART, A. *The Gompertz distribution and maximum likelihood estimation of its parameters: a revision*. Max Planck Institute for Demographic Research, 2012.

MILANI, E. A., TOMAZELLA, V. L. D., DIAS, T. C. M., LOUZADA, F. The generalized time-dependent logistic frailty model: An application to a population-based prospective study of incident cases of lung cancer diagnosed in Northern Ireland. *Brazilian Journal of Probability and Statistics*, v.29, p.132-144, 2015.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing , Vienna, Austria, URL `http://www.R-project.org/`, 2017.

SCHWARZ, G. Estimating the dimension of a model. *The Annals of Statistics*, v.6, p.461-464, 1978.

SINHA, D.; DEY, D. Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*, v.92, p.1195-1212, 1997.

SHARMA, R.; HANAGAL, D. D. Mixture of Gompertz regression model with different frailty distributions. *3^{rd} International Conference on Innovative Approach in Applied Physical, Mathematical/Statistical, Chemical Sciences and Emerging Energy Technology for Sustainable Development*, p.61-68, 2014.

TOMAZELLA, V. L. D.; MARTINS, C. B. ; BERNARDO, J. M. Inference on the univariate frailty model: A Bayesian reference analysis approach. *AIP Conference Proceedings*, v.1073, p.340-347, 2008.

VAUPEL, J. W.; MANTON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, v.16, p.439-454, 1979.

WIENKE, A. *Frailty models in survival analysis*. London: Chapman & Hall/CRC Biostatistics series, 2011.

WILKINSON, W. O. *Lung cancer in Northern Ireland*. Thesis (Master) - Queen's University of Belfast, Belfast, 1995.

Received on 08.06.2017.

Approved after revised on 12.03.2018.

Appendix

In the case of uncensored lifetimes, the log of the likelihood function is given by,

$$\log(L(\alpha,\boldsymbol{\beta},\theta|\boldsymbol{t},\boldsymbol{x}) = -\sum_{i=1}^{n}\frac{\log\left(\frac{\theta\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)}{\theta} - \sum_{i=1}^{n}\log\left(\frac{\theta\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)+\sum_{i=1}^{n}\log\left(e^{\alpha t_i+\boldsymbol{x}_i'\boldsymbol{\beta}}\right). \quad (21)$$

From (21), the first derivatives to the θ, α and β parameters are given respectively by,

$$\frac{\partial \log L(.)}{\partial\theta} = -\sum_{i=1}^{n}\left(\frac{\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha\theta\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)}-\frac{\log\left(\frac{\theta\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)}{\theta^2}\right) - \sum_{i=1}^{n}\frac{\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)},$$

$$\frac{\partial \log L(.)}{\partial\alpha} = -\sum_{i=1}^{n}\frac{\frac{\theta t_i e^{\alpha t_i+\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}-\frac{\theta\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^2}}{\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1}-\sum_{i=1}^{n}\frac{\frac{\theta t_i e^{\alpha t_i+\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}-\frac{\theta\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^2}}{\theta\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)} + \sum_{i=1}^{n}t_i$$

and

$$\frac{\partial \log L(.)}{\partial\beta} = -\sum_{i=1}^{n}\frac{x_i\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)}-\sum_{i=1}^{n}\frac{\theta x_i\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)}+\sum_{i=1}^{n}x_i.$$

In the same way, from (21) we wrote the second derivatives to the θ, α and β parameters are given respectively by,

$$\frac{\partial^2 \log L(.)}{\partial \theta^2} = \sum_{i=1}^{n}\left(-\frac{\left(e^{\alpha t_i}-1\right)^2 e^{2\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^2\theta\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)^2}+\frac{2\log\left(\frac{\theta\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)}{\theta^3}\right.$$
$$\left.-\frac{2\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha\theta^2\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)}\right)+\sum_{i=1}^{n}-\frac{\left(e^{\alpha t_i}-1\right)^2 e^{2\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^2\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)^2},$$

$$\frac{\partial^2 \log L(.)}{\partial \alpha^2} = \sum_{i=1}^{n}\left(\frac{\theta x_i^2\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)}-\frac{\theta^2 x_i^2\left(e^{\alpha t_i}-1\right)^2 e^{2\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^2\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)^2}\right)$$
$$+\sum_{i=1}^{n}\left(\frac{x_i^2\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)}-\frac{\theta x_i^2\left(e^{\alpha t_i}-1\right)^2 e^{2\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^2\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)^2}\right)$$

and

$$\frac{\partial^2 \log L(.)}{\partial \beta^2} = \sum_{i=1}^{n}\left(\frac{\frac{2\theta\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^3}-\frac{2\theta t_i e^{\alpha t_i+\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^2}+\frac{\theta t_i^2 e^{\alpha t_i+\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}}{\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1}\right.$$
$$\left.-\frac{\left(\frac{\theta t_i e^{\alpha t_i+\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}-\frac{\theta\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^2}\right)^2}{\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)^2}\right)$$
$$+\sum_{i=1}^{n}\left(\frac{\frac{2\theta\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^3}-\frac{2\theta t_i e^{\alpha t_i+\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^2}+\frac{\theta t_i^2 e^{\alpha t_i+\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}}{\theta\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)}\right.$$
$$\left.-\frac{\left(\frac{\theta t_i e^{\alpha t_i+\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}-\frac{\theta\left(e^{\alpha t_i}-1\right)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha^2}\right)^2}{\theta\left(\frac{\theta(e^{\alpha t_i}-1)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\alpha}+1\right)^2}\right).$$