UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Non-proportional hazards model with a frailty term: Application with a melanoma dataset

Karen Cristine Ferreira Rosa

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

ICMC USP
SÃO CARLOS

Karen Cristine Ferreira Rosa

Non-proportional hazards model with a frailty term: Application with a melanoma dataset

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master – Professional Masters in Mathematics, Statistics and Computing Applied to Industry. *FINAL VERSION*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
May 2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Cristine Ferreira Rosa, Karen
C788m
Modelo de riscos não proporcionais com um termo de fragilidade: Aplicação em dados de melanoma / Karen Cristine Ferreira Rosa; orientador Francisco Louzada Neto. -- São Carlos, 2021.
59 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Análise de sobrevivência. 2. Distribuição PVF. 3. Fração de cura. 4. Modelo de fragilidade. 5. Modelo de riscos não proporcionais. I. Louzada Neto, Francisco , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Karen Cristine Ferreira Rosa

Modelo de riscos não proporcionais com um termo de fragilidade: Aplicação em dados de melanoma

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestra – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *VERSÃO REVISADA*

Área de Concentração: Matemática, Estatística e Computação

Orientador: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
Maio de 2021

*Este trabalho é dedicado a todas as mulheres que,
muitas vezes sonham em se tornar cientistas.
Em especial a minha mãe, que mesmo sem ter tido muitas oportunidades, sempre nos ensinou o poder da educação.*

ACKNOWLEDGEMENTS

Meus agradecimentos principais são direcionados aos meus pais, Helena e Delcides que nunca deixaram de acreditar em mim, até mesmo quando eu deixei de acreditar. Aos meus irmãos, Jackson e Peterson pelos conselhos e broncas, e algumas brigas também, mas principalmente pelo amor, cuidado e inspiração. Aos meus sobrinhos Duda e Enzo pelos ensinamentos diários, pois é maravilhoso ver o quanto uma criança é capaz de te ensinar até mesmo com um simples sorriso. Ao meu esposo Vinicius Calsavara que esteve ao meu lado nessa caminhada desde o primeiro dia, tutoreando-me, incentivando-me e levantando-me nos momentos de fraqueza. Agradeço também ao meu orientador Prof. Dr. Francisco Louzada Neto, pela oportunidade de realizar esse trabalho.

RESUMO

ROSA, K.C.F. **Modelo de riscos não proporcionais com um termo de fragilidade: Aplicação em dados de melanoma**. 2021. 59 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Em análise de dados de sobrevivência, comumente o tradicional modelo de riscos proporcionais de Cox é ajustado aos dados devido à fácil interpretação das covariáveis sobre a taxa de falha. A principal vantagem deste modelo é a fácil interpretação, a menos que a razão de riscos não variem ao longo do tempo. No entanto, em diversos problemas a suposição de proporcionalidade de uma determinada covariável pode não ser válida, e neste caso, uma abordagem adequada é necessária. Em estudos clínicos é comum uma fração de pacientes não apresentar o evento de interesse (óbito/ recorrência), mesmo se acompanhados por um longo período de tempo, o qual são chamados de imunes ou de fração de curados. Na literatura há diversos modelos de longa duração que contemplam tais situações. Neste trabalho, propomos um modelo de riscos não proporcionais com um termo de fragilidade multiplicativo na função de risco a fim de controlar a heterogeneidade não observável das unidades em estudo com a possibilidade de longa duração. Consideramos uma extensão do modelo log-log generalizado dependente do tempo utilizando a distribuição de fragilidade *Power Variance Function* (PVF) como alternativa para modelar dados de análise de sobrevivência no contexto de riscos não proporcionais na presença ou não de pacientes imunes ao evento de interesse. Estudos de simulações e uma aplicação a dados reais indicam que o modelo proposto pode ser uma ferramenta importante no contexto de riscos não proporcionais.

Palavras-chave: Análise de sobrevivência, Distribuição PVF, Fração de cura, Melanoma, Modelo de fragilidade, Modelo de riscos não proporcionais, Modelo log-log generalizado dependente do tempo.

ABSTRACT

ROSA, K.C.F. **Non-proportional hazards model with a frailty term: Application with a melanoma dataset**. 2021. 59 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

In the modeling of survival data, commonly, the traditional semiparametric Cox regression model is fitted to the dataset due to its ease of interpretation, as long as the hazard rates for two individuals do not vary over time. However, in some situations, the proportionality assumption of the hazards can not be valid. In medical studies, it is expected that a fraction of units do not become susceptible to the event of interest (death or recurrence), even if a sufficiently large time was accompanied, e.g., the so-called long-term survivors. There are several cure rate models available in the literature. Here, we propose the generalized time-dependent complement log-log (CLL) model with a power variance function (PVF) frailty term introduced in the hazard function to control the amount of unobservable heterogeneity in the sample the possibility of long-term survivors. The maximum likelihood estimation procedure reaches the parameter estimation, and we evaluate the performance of the proposed models using Monte Carlo simulation studies. The proposed model's practical relevance is illustrated by applying a dataset on patients diagnosed with skin cancer in the state of São Paulo, Brazil.

Keywords: Frailty model, Generalized time-dependent log-log model, Long-term survivors, Melanoma, Non-proportional hazards model, Power variance function (PVF) distribution, Survival Model.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER 1

INTRODUCTION

In oncology, clinical outcomes are fundamental for all healthcare providers and public policies. Commonly, the researchers' main interest is to estimate the survival rates, as overall survival, cancer-specific survival, or disease-free survival. This information can be obtained based on the cancer type and patient features, such as the gender, age at diagnosis, clinical stage of the disease, education level, type of treatment, and other available information in medical records. In 2018 were expected, approximately $6,000$ new cases of melanoma in Brazil, according to the Brazilian National Institute of Cancer (INCA) (Coordenação de Prevenção e Vigilância, 2017); whereas, according to the International Agency for Research on Cancer (IARC), approximately 7,000 new cases were reported (IARC, available at gco.iarc.fr). Approximately 2,000 deaths per year are attributable to melanoma in Brazil (Coordenação de Prevenção e Vigilância, 2017; ERVIK *et al.*, 2020).

In the context of survival analysis, the traditional semiparametric Cox regression model (COX, 1972) is often fitted to censored survival data to evaluate the effect of covariates in the hazard rate, which assumes that it is constant over time. In practice, the covariate effects may change over time, and the Cox model may not be adequate. It is common in a clinical study where prognostic factors, such as treatment, disappear with time. As mentioned in Calsavara *et al.* (2020), some types of tumors may respond well to chemotherapy/radiotherapy initially, but the cancer cells may develop some tolerance to the treatment through genetic mechanisms, resulting in loss of the treatment effect over time. A problem to fit a Cox regression model can be inadequate and wrong conclusions can be made. As observed in Schemper (1992), the Cox model has undoubtedly been used in many problems, proportionality assumptions are violated, and consequences for the results.

In practice, the Cox regression model is commonly fitted to the dataset, and the proportionality assumption is assessed through Schoenfeld residuals (SCHOENFELD, 1982; PETTITT; DAUD, 1990; GRAMBSCH; THERNEAU, 1994). Klein and Moeschberger (2003) suggest plotting the log of the cumulative hazard functions against time and checking for parallelism. In the literature, there are several graphical methods for the assessment of this assumption. Eight graphical methods for the detection of assumption violations have been proposed by Hess (1995). If departures of the assumption are detected, we can consider some possible workarounds, such as the redefinition of covariates, model stratification by a covariate with a non-proportional hazard, fitting a non-proportional hazard model, and so on.

There are several approaches to deal with non-proportional hazards, as the nonparametric accelerated failure time model proposed by Prentice (1978) and Kalbfleisch and Prentice (2011), Etezadi-Amoli and Ciampi (1987) considered the hybrid hazard model. In contrast, Louzada-Neto (1997), Louzada-Neto (1999) proposed the extension of hybrid hazard models, the generalized time-dependent logistic (GTDL) model proposed by Mackenzie (1996) and extended by Milani *et al.* (2015) for the gamma frailty model approach. Recently, Calsavara *et al.* (2020) extended the GTDL model considering a power variance function (PVF) frailty model and incorporating covariate in the effect term. These models have been applied successfully to problems in which all subjects are susceptible to the event of interest. However, in several situations, some subjects will not fail during the follow-up because the units are long-term survivors or immune to the event of interest. In this sense, the long-term survival models consider such situations and have been extensively studied by several authors. The mixture model proposed by Boag (1949) and modified by Berkson and Gage (1952) is the most popular model, where the population survival function is $S(t) = p + (1-p)S_0(t)$, such that $p \in (0,1)$ is the the long-term survivors, and $S_0(t)$ is a proper survival function for susceptible patients. The exponential and Weibull distributions are common choices for $S_0(t)$. Nevertheless, other distribution can be considered.

The use of traditional models in survival analysis and the long-term survival models can be extended to capture the effects of unobserved covariate that were not incorporated in the model, such as genetic factors, environmental or information that was not considered in planning. Hougaard (1991) showed the advantages of considering two sources of heterogeneity (observable and unobservable) in a model. A way to quantify the unobservable heterogeneity is employing frailty models, in which a random effect is considered (multiplicatively or additive form) in the hazard function to represent the information that cannot be or has not been observed. The random effect also allows the assessment of covariate effects that were not considered. If an important covariate is omitted from the model, the amount of unobservable heterogeneity will increase, affecting the model parameters' inferences. To include a random effect can help to alleviate this problem (CALSAVARA *et al.*, 2020). Frailty models have been studied by several authors Clayton (1978), Vaupel, Manton and Stallard (1979), Andersen *et al.* (2012), Hougaard (1995), Sinha and Dey (1997), Oakes (1982). Other authors as Aalen (1988), Hougaard, Myglegaard and Borch-Johnsen (1994), Price and Manatunga (2001), Peng, Taylor and Yu (2007), Yu and Peng (2008), Calsavara, Tomazella and Fogo (2013), Calsavara *et al.* (2017) have considered cure rate models with frailty terms. Recently, Calsavara *et al.* (2020) proposed a non-proportional hazards frailty model with the possibility of long-term survivors in patients diagnosed with melanoma in the state of São Paulo, Brazil.

Another way to modeling the long-term survivors is through defective models recently proposed by Balka, Desmond and McNicholas (2009), Balka, Desmond and McNicholas (2011), Rocha *et al.* (2016), Rocha *et al.* (2017a), Rocha *et al.* (2017b), Scudilio *et al.* (2019) and Calsavara *et al.* (2019a), Calsavara *et al.* (2019b). Defective distributions are obtained from standard distributions by changing the domain of the latter's parameters so that their survival functions are limited to $p \in (0,1)$ (CALSAVARA *et al.*, 2019b). An advantage of this model is the ability to accommodate or not cure fractions depending only on shape parameter value, which is very interesting in terms of the flexibility and capacity to model different types of data sets.

Here, we propose a different way to model lifetime data under non-proportional hazards and possibly a cure fraction of the population. In addition, we included a frailty term on the baseline hazard function to deal with possible heterogeneity due to unobserved covariates. In the next section, we show the motivation of the proposed model using a real cancer dataset.

1.1 Motivation

The proposed model is motivated by a real medical dataset. It is part of a study of skin cancer in 6752 patients diagnosed melanoma in the state of São Paulo, Brazil. Melanoma is one of the best known by the population, but skin carcinomas are more incidents than melanoma. The survival of patients with melanoma is worse due to its potential for metastatic dissemination. In general, patients in the clinical stages I or II are treated with surgery, and most of them will be alive after ten years of follow-up, while patients with clinical stages advanced other therapies are conducted, and its prognosis is worse due to its potential for metastatic dissemination.

The patients were enrolled in the study from 2000 to 2014, with follow-up conducted until 2018. They were followed after the diagnosis, and the death due to cancer was defined as the event of interest. Patients with lost follow-up or died due to other causes in the follow-up period were characterized as right-censored observations. All records were provided by the São Paulo Oncocenter Foundation (FOSP), and they can be downloaded in http://www.fosp.saude.sp.gov.br. The hospital cancer registry (RHC/FOSP) started its activities in 2000, intending to register cancer cases treated in the state. Currently, 77 hospital cancer registries are active, and every three months, the records send the datasets. The FOSP is a public institution connected to the State Health Secretariat, which assists in preparing and implementing healthcare policies in Oncology. As mentioned by Andrade *et al.* (2012), these policies serve as an instrument for oncology hospitals to prepare their protocols and improve care practices.

The melanoma data set considered in this project was also initially studied by Calsavara *et al.* (2020), where they evaluated only the effect of surgery covariate in a lifetime using a non-proportional hazards model with a frailty term. In our study, the goal was to assess the effect of surgery and other observed covariates available in the record, such as gender and age at diagnostic on specific survival. A total of 414 patients were removed from the sample due to missing values on the covariates observed, leaving 6752 patients in the study. Of the 6752 patients, 5981 (88.6%) underwent surgery and 771 (11.4%) did not, 3417 (50.6%) were female and 3335 (49.4%) male, 2201 (32.6%) were younger ($\leq$ 50 years-old) and 4551 (67.4%) older ($>$ 50 years-old). A total of 1914 (28.3%) events occurred during the follow-up period: Approximately 18.54 years was the maximum observation time, and the median follow-up time was 5.19 years.

In the literature, the overall melanoma-specific survival after ten years may vary from 24% to 88% (GERSHENWALD *et al.*, 2017). Figure 1 shows the estimated overall melanoma-specific survival obtained by Kaplan-Meier estimator (KAPLAN; MEIER, 1958) for the melanoma dataset. The 5-, 10-, 15- and 18-year specific survival rates were 0.706, 0.629, 0.598 and 0.591, respectively. We observe the presence of long-term survivors, as expected.

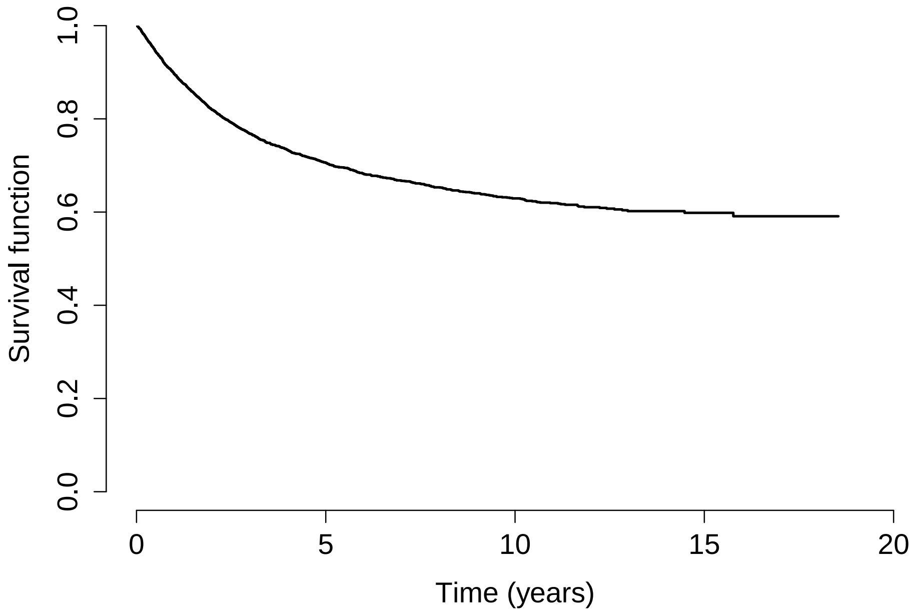

Figure 1 – Estimated melanoma-specific survival obtained by Kaplan-Meier estimator for melanoma dataset.

We provide in Figure 2 the estimated survival function for each observed covariate. According to the estimated survival curves, patients in the surgery group have a better prognosis, as expected, since most of the patients were in the early stage of the disease and these patients are normally treated with surgery; better survival rates are associated with young female patients. In addition, there is evidence of long-term survivors for each observed covariate.

In melanoma skin cancer, there are several relevant information about the patients that

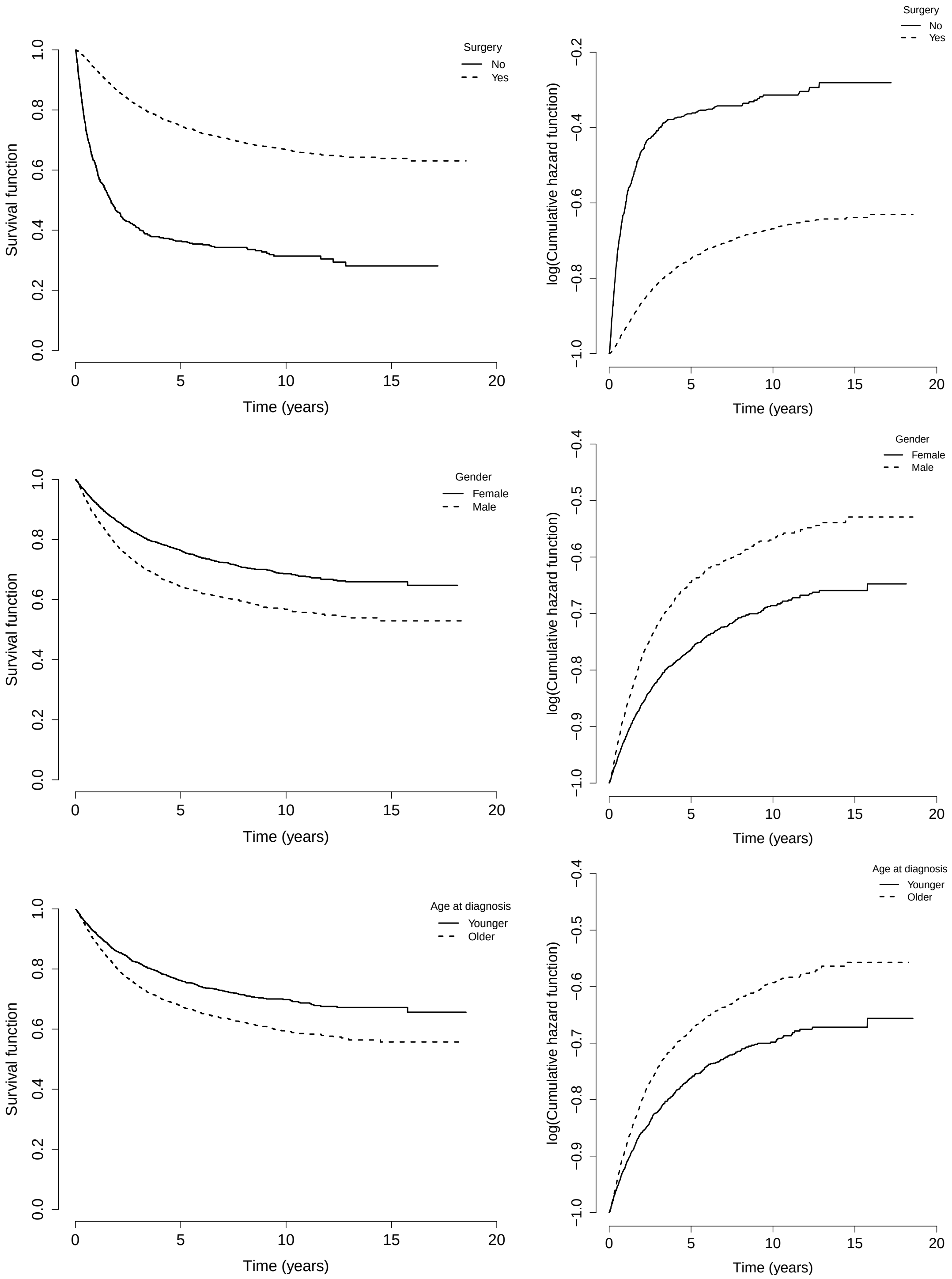

Figure 2 – Estimated survival curve obtained via Kaplan-Meier estimator for melanoma dataset (left panel) and plot of log cumulative baseline hazard rates versus time on study (right panel).

could be considered in the analysis, as Breslow thickness[1], ulceration[2] and Mitotic rate[3], as well

[1] Breslow thickness is the single most important prognostic factor for clinically localized primary melanoma. Breslow thickness is measured from the top of the granular layer of the epidermis (or, if the surface is ulcerated, from the base of the ulcer) to the deepest invasive cell across the broad base of the tumor (dermal/subcutaneous) as described by Breslow.

[2] Ulceration is an integral component of the AJCC/UICC staging system and an independent predictor

as the environment and genetic factors. However, due to several reasons, significant covariates were not observed or can not be observed.

Figure 2 also shows a plot of log cumulative baseline hazard rates against time (follow-up period) for the surgery, gender, and age at diagnosis. According to Klein and Moeschberger (KLEIN; MOESCHBERGER, 2003), if the proportionality assumption holds, then these curves should be approximately parallel, with constant vertical separation between them. The plots suggest non-proportional hazards for the surgery covariate. In particular, the proportionality is questionable before 5 years, as also observed in Calsavara *et al.* (2020). Thus, to fit the traditional Cox model to this dataset can not be adequate.

In this sense, given the structure of the dataset is necessary adequate modeling to deal with long-term survivors and non-proportional hazards, as well as incorporating unobserved information in the modeling, once those significant covariates were not observed, such as Breslow thickness, ulceration, and Mitotic rate (BERTOLLI *et al.*, 2019; FONSECA *et al.*, 2020).

1.2 Objectives

The main goal is to evaluate the effect of the surgery in the lifetime adjusted by age at diagnosis and gender; quantify the amount of unobserved heterogeneity due to lack of relevant clinical information, and estimate the long-term survivors.

The traditional semiparametric Cox regression model and the long-term survival models are essential models in the survival analysis, and several variations of these models have been proposed in the literature to lead with this real problem. However, our strategy is to consider the generalized time-dependent complement log-log (CLL) model by including a scale parameter (λ), as well as a power variance function frailty term in the modeling, which is an extension of the model proposed by Milani, Diniz and Tomazella (2014).

The dissertation is organized as follows: Chapter 2 a brief review of the concepts of survival analysis is presented, the Cox regression model, and the idea of frailty model. Furthermore, the non-proportional hazard model focuses on the generalized time-dependent complement log-log model, and the maximum likelihood estimation is presented. In Chapter 3 is presented the new non-proportional hazard model with a PVF frailty term, and its properties are discussed. To finalize this chapter, a simulation study is presented to analyze the asymptotic properties of maximum likelihood estimators under different sample size scenarios. In Chapter 4 we apply the proposed model to the real melanoma cancer dataset. Finally, in Chapter 5 we discuss the conclusions of this dissertation and some proposals for future work.

of outcome in patients with clinically localized primary cutaneous melanoma.

[3] Multiple studies indicate that mitotic count is an important prognostic factor for localized primary melanoma since it represents tumor cells division.

CHAPTER 2

BACKGROUND

2.1 Introduction

This chapter aims to briefly present background about survival analysis and the traditional semiparametric Cox regression model and its frailty model version. Furthermore, the non-proportional hazard model focuses on the generalized time-dependent complement log-log model, and the maximum likelihood estimation is presented.

2.2 Survival analysis

The main interest in survival analysis is to estimate the time until an event of interest. For instance, in clinical studies, it is common for researchers to have an interest in estimate the lifetime of patients in different types of treatments, the time until the disease recurrence, and so on. However, in survival data is expected the presence of incomplete observations, the so-called censored observations. Indeed, a proportion of units are expected, which are not susceptible to the event of interest (death or recurrence). The presence of censoring in the sample impossible to apply standard statistical for analyzing such data. Therefore, appropriate techniques are needed that take into account partial information. There are some kinds of censoring, as type I and II censoring, random censoring, left or interval censoring, etc.

According to Klein and Moeschberger (2003), the type I censoring (or right-censoring) occurs when the event is observed only if it occurs before some prespecified time. Type II censoring occurs when a study with n units continues until the failure of the first r individuals, where r is some predetermined integer ($r < n$). The random censored observation, unlike the others, occurs when the unit leaves the study without having experienced the event of interest. In practice, the random censoring observation is a more typical case. Interval censoring is a more general type of censoring that occurs when the lifetime is only known to occur within an interval $(L, R]$, where L denotes the left endpoint and R for the right endpoint. Such interval censoring occurs when patients in a clinical trial or longitudinal study have a periodic follow-up, and the patient's event time is only known to fall in an interval.

In survival data each observation is denoted by (t_i, δ_i), where t_i denotes the time until the failure or censoring and δ_i is the censoring indicator variable, that is, $\delta_i = 0$ if the observed time is censored and $\delta_i = 1$ otherwise, $i = 1, \ldots, n$. Let $T > 0$ be a random variable representing the time until the occurrence of the event of interest with density function $f(t)$. The density function is defined as the limit of the probability of a subject fails in the interval of time $[t; t + \Delta t]$ as follows

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t},$$

and its cumulative function is

$$F(t) = P(T \leq t) = \int_0^t f(u)du.$$

The function of major interest in survival analysis is the survival function. It represents the probability of an individual survival at least until the time t and it is given by

$$S(t) = P(T \geq t) = \int_t^\infty f(u)du = 1 - F(t).$$

The survival function properties are: $S(t)$ is nonincreasing function; $S(0) = 1$ and $\lim_{t \to \infty} S(t) = 0$. When the last property is satisfied the survival function is said proper survival function.

Another important function in survival analysis is the hazard function and it provides the instant rate of fail. This function represents the chance of a subject will fail in the time $t + \Delta t$, with $t \to 0$ and it is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

The hazard function can have several shapes, and the most studied cases are: constant, decrease, increase, unimodal, and bathtub shaped.

Its cumulative hazard function is given by

$$H(t) \quad = \quad \int_0^t h(u)du.$$

Some useful relations between them are

$$\begin{aligned} f(t) &= -\frac{dS(t)}{dt} \\ S(t) &= \exp\{-H(t)\} \\ h(t) &= \frac{f(t)}{S(t)}. \end{aligned}$$

2.3 Kaplan-Meier estimator

In the literature there are some estimators of survival function. The most important non-parametric estimator and used in the practice is the Kaplan-Meier estimator (KAPLAN; MEIER, 1958) and it is given by

$$\widehat{S}(t) = \prod_{j:\, t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:\, t_j<t} \left(1 - \frac{d_j}{n_j}\right), \tag{2.1}$$

where $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ are the k distinct and ordered fail times; d_j the number of fails in $t_{(j)}$, $j = 1, \ldots, k$; n_j represents the number of units at risk in time $t_{(j)}$, i.e., subjects who have not failed or were not censored until the moment instantly previous to $t_{(j)}$.

As mentioned by Klein and Moeschberger (2003), the Kaplan-Meier estimator is a step function with jumps at the observed event times. The size of these jumps depends not only on the number of events observed at each event time t_i but also on the pattern of the censored observations before t_i. According to Kaplan and Meier (1958) the estimator in (2.1) is the maximum likelihood estimator of $S(t)$. In this sense, the Kaplan-Meier curve is widely applied to verify the goodness-of-fit of the proposed parametric survival models.

2.4 Cox regression model

In the modeling of survival data, commonly, the traditional semiparametric Cox regression model (COX, 1972) is fitted to dataset in order to evaluate the effect of covariates $\mathbf{x}^\top = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ in the lifetime. This model is often considered due to its ease of interpretation, as long as the hazard rates for two individuals do not vary over time. The Cox regression model assumes that.

$$h(t \mid \mathbf{x}) = h_0(t) g(\mathbf{x}^\top \beta), \tag{2.2}$$

where $h_0(t)$ is an arbitrary (non-negative) baseline hazard function, $g(\mathbf{x}^\top \beta)$ is a non-negative function with $g(0) = 1$, often taken as $\exp\{\mathbf{x}^\top \beta\}$, and $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is an unknown p-dimensional parameter.

In general, estimates of β are obtained by the partial likelihood method, but it will not be addressed here. We refer the interested readers to Klein and Moeschberger (2003).

The basic assumption for its use is that failure rates, of any two individuals, are constant over time, i.e, the ratio of the hazard function (2.2) for two individuals, i and j, with $i \neq j$ does not depend of the time, as follows

$$\tau(t; \mathbf{x}_i, \mathbf{x}_j) = \frac{h_0(t)\exp\{\mathbf{x}_i^\top \beta\}}{h_0(t)\exp\{\mathbf{x}_j^\top \beta\}} = \exp\{\beta(\mathbf{x}_i^\top - \mathbf{x}_j^\top)\}.$$

However, there are situations where the covariate effects may change over time, and the traditional Cox regression model may not be adequate. In practice, the covariate effects may change over time. For instance, in a clinical study, prognostic factors, such as treatment may disappear with time. As mentioned in Calsavara *et al.* (2020), some types of tumors may respond well to chemotherapy/radiotherapy initially, but the cancer cells may develop some tolerance to the treatment through genetic mechanisms, resulting in loss of the treatment effect over time. A problem to fit a Cox regression model can be inadequate and wrong conclusions can be made. As observed in Schemper (1992), the Cox model has undoubtedly been used in many problems in which proportionality assumptions are violated, with consequences for the results.

In practice, one usually fits a Cox proportional hazards model and assesses the proportionality assumption based on the so-called Schoenfeld residuals (SCHOENFELD, 1982; PETTITT; DAUD, 1990; GRAMBSCH; THERNEAU, 1994). Klein and Moeschberger (2003) suggest plotting the log of the cumulative hazard functions against time and checking for parallelism. In the literature, there are several graphical methods for the assessment of this assumption. If departures from the assumption are detected, we can consider possible workarounds, such as the redefinition of covariates, model stratification by a covariate with a non-proportional hazard, fitting a non-proportional hazard model, and so on.

2.5 Frailty model

The concept of frailty provides a convenient way of introducing unobserved heterogeneity and associations into models for survival data. In the univariate time scenario, the role of frailty is to measure a possible heterogeneity to identify the influence of covariates that were not incorporated in the model, such as genetic factors, environmental or information that was not considered in planning (CALSAVARA *et al.*, 2020).

The multiplicative frailty model is an extent of Cox proportional hazards model, where the individual hazard depends of an unobservable random variable V, called frailty, which acts multiplicatively on the baseline hazard function. The hazard function of the ith individual with the frailty term v_i multiplicative is given by

$$h_i(t;v_i,\mathbf{x}_i) = v_i h_0(t)\exp\{\mathbf{x}_i^\top \beta\}. \tag{2.3}$$

The conditional hazard function (2.3) is smaller than baseline for $v_i < 1$; greater than the baseline hazard function for $v_i > 1$ and when $v_i = 1$ the frailty model reduces to the Cox regression model (2.2). Due to how the random effect acts on the hazard function, natural frailty distribution candidates are supposed to be non-negative, continuous, and time-independent. The frailty distribution widely used in practice is the gamma distribution with mean 1 and variance θ, as it permits easy algebraic treatment. However, other choices can be considered, as log-normal, positive stable, power variance function distributions.

For simplicity let us consider the model (2.3) without the presence of the covariates.

Thus, the survival function of an individual conditional on the frailty $V = v$ is

$$S(t;v_i) = S_0(t)^{v_i},$$

where $S_0(\cdot)$ denotes the baseline survival function.

In order to eliminate the unobserved quantities, the random effect can be integrated out. Thus, marginal survival function is given by

$$S(t) = \mathbb{E}_V\left[S(t;v)\right] = \int_0^\infty S(t;v) f_v(v) dv = \mathscr{L}_v\left[-\log S_0(t)\right],$$

where $f_v(\cdot)$ is the probability density of the corresponding frailty distribution, $S_0(\cdot)$ is the baseline survival function, and $\mathscr{L}_v[\cdot]$ denotes the Laplace transform of frailty distribution.

2.6 Non-proportional hazard model

In the literature, there are several approaches to lead with non-proportional hazards and long-term survivors. Here we focus on a non-proportional hazards model proposed by Milani, Diniz and Tomazella (2014) which also allows a proportion of units non-susceptible to the event of interest without requiring an extra parameter, which is an advantage over traditional cure rate models.

2.6.1 Complementary log-log hazard model

Let $T > 0$ be a random variable representing the failure time and $h_0(t)$ the instantaneous failure rate or baseline hazard function. According to Milani et al. (MILANI; DINIZ; TOMAZELLA, 2014), the complementary log-log hazard function (CLL) is given by

$$h_0(t;\mathbf{x}) = \exp\{-\exp(\alpha t + \mathbf{x}^\top \beta)\}, \tag{2.4}$$

where α is a measure of the time effect, $\mathbf{x}^\top = (1, \mathbf{x}_1, \ldots, \mathbf{x}_p)$, and $\beta^\top$=$(\beta_0, \beta_1, \ldots, \beta_p)$ are the sets of covariates and their regression coefficients, respectively.

The cumulative hazard function $H(t;\mathbf{x})$ and survival function $S(t;\mathbf{x})$ are, respectively, as follows

$$H_0(t;\mathbf{x}) = \int_0^t \exp\{-\exp(\alpha y + \mathbf{x}^\top \beta)\} dy$$

and

$$S_0(t;\mathbf{x}) = \exp\left(-\int_0^t \exp\left\{-\exp(\alpha y + \mathbf{x}^\top \beta)\right\} dy\right). \tag{2.5}$$

The ratio of the hazard function for two individuals, i and j, with $i \neq j$ where $i, j = 1, \ldots, n$, with different covariate vectors is given by

$$\begin{aligned}\tau(t;\mathbf{x}_i,\mathbf{x}_j) = \frac{h_0(t;\mathbf{x}_i)}{h_0(t;\mathbf{x}_j)} &= \frac{\exp\{-\exp(\alpha t + \mathbf{x}_i^\top \beta)\}}{\exp\{-\exp(\alpha t + \mathbf{x}_j^\top \beta)\}} \\ &= \exp\{-\exp(\alpha t)[\exp(\mathbf{x}_i^\top \beta) - \exp(\mathbf{x}_j^\top \beta)]\}. \end{aligned} \tag{2.6}$$

The time effect does not disappear in (2.6). Consequently, the non-proportionality becomes evident. Note that if $\alpha = 0$, the model is a proportional hazards model.

As mentioned by Milani et al. (MILANI; DINIZ; TOMAZELLA, 2014), the behavior of the hazard function (2.4) depends on the value of α. When $\alpha > 0$, it decreases; when $\alpha < 0$ it is increasing, and when $\alpha = 0$, the hazard function is constant over time. Due to the hazard function's shape, the model (2.4) is indicated for the modeling phenomenon with monotone failure rates.

The survival function is proper for $\alpha \leq 0$, but when time effect α is positive, the CLL model naturally becomes an improper distribution, which is useful for the modeling of survival data in the presence of a surviving fraction. The long-term survivors is calculated as the limit of the survival function (2.5) when $\alpha > 0$, given by

$$p(\mathbf{x}) = \lim_{t\to\infty} S(t;\mathbf{x}) \quad = \quad \lim_{t\to\infty} \exp\left(-\int_0^t \exp\left\{-\exp(\alpha y + \mathbf{x}^\top \beta)\right\} dy\right) \in (0,1). \qquad (2.7)$$

An advantage of the CLL model over traditional cure rate models is that it does not assume the existence of the long-term survivors and neither requires an extra parameter. Moreover, the CLL model has an inconvenient constraint on the hazard function imposed by the $0 \leq h_0(t;\mathbf{x}) \leq 1$, for all $t > 0$.

Figure 3 plots the baseline hazard and survival functions for different parameter values for the CLL model considering a group variable as the covariate. As previously mentioned the hazard function is constraint on unit interval.

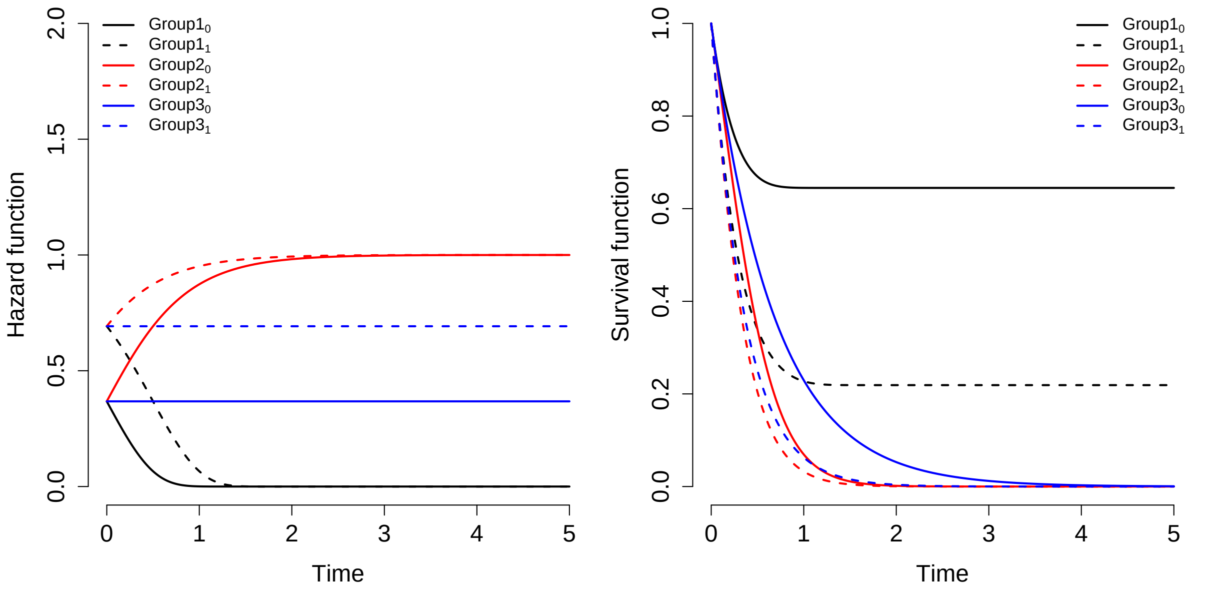

Figure 3 – Baseline hazard (left panel) and survival (right panel) functions from the extended CLL model. The parameter values used are: Group1, $\alpha = 2$, $\beta_0 = 0$, and $\beta_1 = -1$; Group2, $\alpha = -2$, $\beta_0 = 0$, and $\beta_1 = -1$; and Group3, $\alpha = 0$, $\beta_0 = 0$, and $\beta_1 = -1$. The subscript numerals indicate the values of the fixed covariates.

2.7 Maximum likelihood estimation

There are several non-parametric approaches in survival analysis to estimate the survival function, cumulative hazard function, among other interest functions. However, our focus here is to fit parametric models to the observed data. The maximum likelihood estimation is widely used in statistical to estimate parameters of statistical models. The maximum likelihood method allows to incorporate the incomplete observations commonly observed on survival datasets, and it has excellent properties for large samples. The contribution of censored observations to the likelihood function is given by the survival function, while the complete observations contribute to the density function.

Let us consider the situation when the time to event is not completely observed and it is subject to right censoring. Let δ_i the censoring indicator variable, that is, $\delta_i = 0$ if the observed time is censored and $\delta_i = 1$ otherwise, $i = 1, \ldots, n$. The observed dataset is $\mathbf{D} = (\mathbf{t}, \delta, \mathbf{X})$, where $\mathbf{t} = (t_1, \ldots, t_n)^\top$ are the observed lifetimes, $\delta = (\delta_1, \ldots, \delta_n)^\top$ are the censoring indicators, and $\mathbf{X}$ is a matrix containing the covariate information. Consider that T_is are independent and identically distributed random variables with survival, hazard and density functions specified, respectively, by $S(\cdot; \vartheta, x)$, $h(\cdot; \vartheta, x)$ and $f(\cdot; \vartheta, x)$, where ϑ denotes a vector of unknown parameters. We suppose that T is independent of the censoring time. Under non-informative censoring the likelihood function of ϑ is

$$\begin{aligned} L(\vartheta; \mathbf{D}) &\propto \prod_{i=1}^{n} [f(t_i; \vartheta, x_i)]^{\delta_i} [S(t_i; \vartheta, x_i)]^{1-\delta_i} \\ &\propto \prod_{i=1}^{n} h(t_i; \vartheta, x_i)^{\delta_i} S(t_i; \vartheta, x_i). \end{aligned}$$

The corresponding log-likelihood function, $\ell(\vartheta) = \log L(\vartheta; \mathbf{D})$, is given by

$$\ell(\vartheta) \propto \sum_{i=1}^{n} \delta_i \log f(t_i; \vartheta, x_i) + \sum_{i=1}^{n} (1 - \delta_i) \log S(t_i; \vartheta, x_i).$$

The maximum likelihood estimator is the value of ϑ that maximizes $L(\vartheta)$ or equivalently its log-likelihood function $\ell(\vartheta)$. The estimators are found by solving the system of equations

$$U(\vartheta) = \frac{\partial \ell(\vartheta)}{\partial \vartheta} = \mathbf{0}.$$

Commonly, due to the complexity of the equations, the maximum likelihood estimator does not have a closed expression. Thus, it is necessary to use numeric methods. There are many routines available for numerical maximization in the literature. We used the *optim* routine in the R software (R Core Team, 2020) for numerical maximization with the L-BFGS-B optimization method.

The asymptotic properties of maximum likelihood estimators are needed to obtain the confidence intervals and to test hypotheses about the model parameters. Under certain regularity conditions, $\widehat{\vartheta}$ has asymptotic multivariate normal distribution with mean ϑ and variance and covariance matrix $\Sigma(\widehat{\vartheta})$, which is estimated by

$$\widehat{\Sigma}(\widehat{\vartheta}) = \left\{ -\left.\frac{\partial^2 \ell(\vartheta;\mathbf{D})}{\partial\vartheta\partial\vartheta^\top}\right|_{\vartheta=\widehat{\vartheta}} \right\}.$$

Thus, an approximate $100(1-\alpha^*)\%$ confidence interval for ϑ_i is $\left(\widehat{\vartheta}_i - z_{\alpha^*/2}\sqrt{\widehat{\Sigma}^{ii}}, \widehat{\vartheta}_i + z_{\alpha^*/2}\sqrt{\widehat{\Sigma}^{ii}}\right)$, where $\widehat{\Sigma}^{ii}$ denotes the ith diagonal element of the inverse of $\widehat{\Sigma}$ and z_{α^*} denotes the $100(1-\alpha^*)$ percentile of the standard normal random variable.

In the CLL model, the long-term survivors $p(\mathbf{x})$ is calculated as a function of the estimated parameters. Due the complexity of integral in (2.7), numerical integration is necessary, as adaptive quadrature of functions of one variable (GANDER; GAUTSCHI, 2000). We considered the *integrate* function available in R software to approximate the integral once it has no analytical solution. In addition, to estimate the standard error and confidence interval of $p(\mathbf{x})$, the non-parametric bootstrap technique can be used. For more details, please see Davison and Hinkley (1997).

CHAPTER

3

EXTENDED COMPLEMENTARY LOG-LOG HAZARD MODEL WITH A FRAILTY TERM

3.1 Introduction

This chapter aims to propose a new family of non-proportional hazard survival models with a frailty term and the possibility of long-term survivors. This new family is obtained by adding an extra parameter (λ) in the CLL model and a frailty term in the hazard function to quantify the amount of unobserved heterogeneity.

3.2 Extended complementary log-log hazard model

As previously mentioned in Section 2.6, the complementary log-log hazard function (CLL) is given by

$$h_0(t;\mathbf{x}_1) = \exp\{-\exp(\alpha t + \mathbf{x}_1^\top \beta)\}, \tag{3.1}$$

where α is a measure of the time effect, $\mathbf{x}_1^\top = (1, \mathbf{x}_{1_1}, \ldots, \mathbf{x}_{1_p})$, and $\beta^\top$=$(\beta_0, \beta_1, \ldots, \beta_p)$ are the sets of covariates and their regression coefficients, respectively.

As previously mentioned, the CLL model does not assume the existence of the long-term survivors, and neither requires an extra parameter to estimate the cure rate. Moreover, it has an inconvenient constraint on the hazard function $0 \leq h_0(t;\mathbf{x}_1) \leq 1$, for all $t > 0$.

Due the limitation on the hazard function on unit interval, we propose to incorporate a scale parameter $\lambda > 0$ into the hazard function (3.1) in order to became the model more realistic. Thus, the evolution of the CLL model to the extended CLL model is due to the inclusion multiplicative of λ in (3.1), which the hazard function does not limited in the interval $[0,1]$. Therefore, the extended complementary log-log model (or extended CLL model) is given by

$$h_0(t;\mathbf{x}_1) = \lambda \exp\{-\exp(\alpha t + \mathbf{x}_1^\top \beta)\}, \tag{3.2}$$

where $\lambda > 0$ is the scale parameter, α is a measure of the time effect, $\mathbf{x}_1^\top = (1, \mathbf{x}_{1_1}, \ldots, \mathbf{x}_{1_p})$, and $\beta^\top$=$(\beta_0, \beta_1, \ldots, \beta_p)$ are the sets of covariates and their regression coefficients, respectively.

Its corresponding cumulative hazard function $H(t;\mathbf{x}_1)$ and survival function $S(t;\mathbf{x}_1)$ are, respectively, as follows

$$H_0(t;\mathbf{x}_1) = \int_0^t \lambda \exp\{-\exp(\alpha y + \mathbf{x}_1^\top \beta)\} dy \tag{3.3}$$

and

$$S_0(t;\mathbf{x}_1) = \exp\left(-\int_0^t \lambda \exp\left\{-\exp(\alpha y + \mathbf{x}_1^\top \beta)\right\} dy\right). \tag{3.4}$$

As does the CLL model, the ratio of the hazard function in the extended CLL model for two individuals is also time-dependent, as follows

$$\begin{aligned} \tau(t;\mathbf{x}_{1_i},\mathbf{x}_{1_j}) = \frac{h_0(t;\mathbf{x}_{1_i})}{h_0(t;\mathbf{x}_{1_j})} &= \frac{\lambda \exp\{-\exp(\alpha t + \mathbf{x}_{1_i}^\top \beta)\}}{\lambda \exp\{-\exp(\alpha t + \mathbf{x}_{1_j}^\top \beta)\}} \\ &= \exp\{-\exp(\alpha t)[\exp(\mathbf{x}_{1_i}^\top \beta) - \exp(\mathbf{x}_{1_j}^\top \beta)]\}, \end{aligned} \tag{3.5}$$

for $i \neq j$ where $i, j = 1, \ldots, n$, with different covariate vectors.

The survival function is also proper for $\alpha \leq 0$ and improper when time effect α is positive. The long-term survivors, when $\alpha > 0$ is

$$p(\mathbf{x}_1) = \lim_{t\to\infty} S(t;\mathbf{x}_1) = \lim_{t\to\infty} \exp\left(-\lambda \int_0^t \exp\left\{-\exp(\alpha y + \mathbf{x}_1^\top \beta)\right\} dy\right) \in (0,1). \tag{3.6}$$

The shape of the hazard function (3.2) takes the same forms of the CLL model (3.1), as previously mentioned. However, it is not limited to unit intervals. Also note that the usual CLL model (3.1) is obtained if $\lambda = 1$. Figure 4 plots the baseline hazard and survival functions for different parameter values for the extended CLL model considering a group variable as the covariate.

3.3 Extended complementary log-log frailty hazard model

The multiplicative frailty model is an extension of the proportional hazards model introduced by Cox (COX, 1972), where the unit's hazard function depends on a non-negative unobservable random variable V, which acts multiplicatively on the baseline hazard function. From the extended CLL model (3.2), the hazard function of the ith individual with the frailty term v_i is given by

$$h_i(t;v_i,\mathbf{x}_{1_i}) = v_i h_0(t;\mathbf{x}_{1_i}) = v_i \lambda \exp\{-\exp(\alpha t_i + \mathbf{x}_{1_i}^\top \beta)\}. \tag{3.7}$$

The conditional hazard function (3.7) is smaller than baseline for $v_i < 1$; greater than the baseline hazard function for $v_i > 1$ and when $v_i = 1$ the frailty model reduces to the CLL

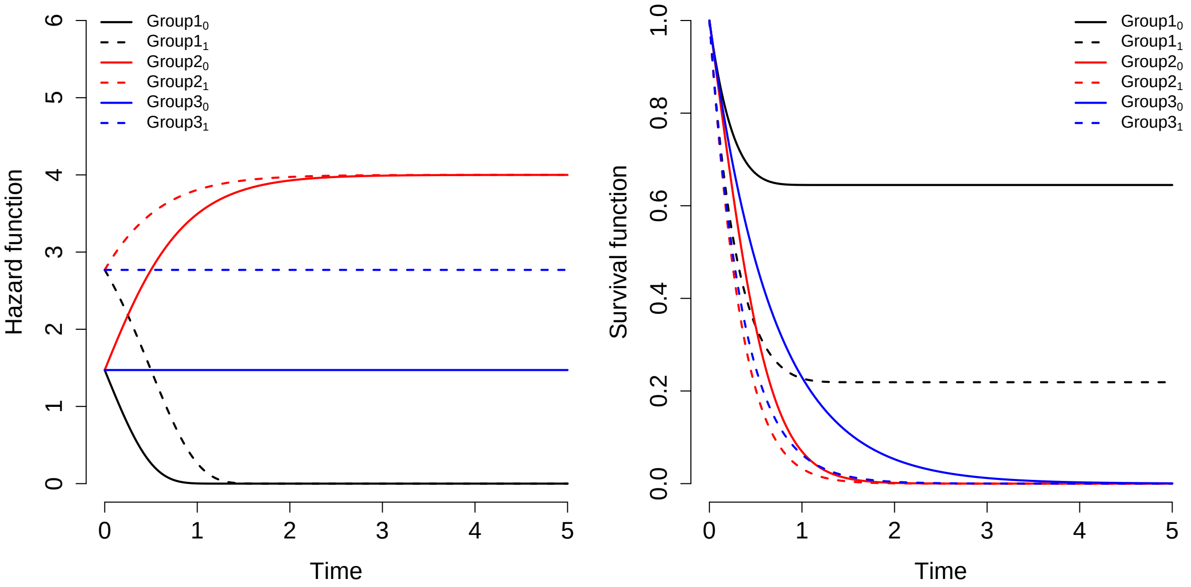

Figure 4 – Baseline hazard (left panel) and survival (right panel) functions from the extended CLL model. The parameter values used are: Group1, $\alpha = 2$, $\lambda = 4$, $\beta_0 = 0$, and $\beta_1 = -1$; Group2, $\alpha = -2$, $\lambda = 4$, $\beta_0 = 0$, and $\beta_1 = -1$; and Group3, $\alpha = 0$, $\lambda = 4$, $\beta_0 = 0$, and $\beta_1 = -1$. The subscript numerals indicate the values of the fixed covariates. (For interpretation of the references to color in this figure legend, the reader is referred to the online version of this article.)

model (3.1). Due to how the random effect acts on the hazard function, natural frailty distribution candidates are supposed to be non-negative, continuous, and time-independent. The gamma and inverse Gaussian distributions were considered in Milani et al. (MILANI; DINIZ; TOMAZELLA, 2014) with mean 1 and variance θ for the random effect. However, other choices can be considered, as log-normal, positive stable, power variance function distributions.

In this work, we consider the family of power variance function (PVF) distributions, as it presents as a particular case the gamma, inverse Gaussian, and positive stable distributions. The PVF distribution was suggested by Tweedie (TWEEDIE, 1984) and derived independently by Hougaard (HOUGAARD, 1986). Let V be a random variable following a PVF distribution with parameters μ, ψ and γ with density function written as (WIENKE, 2011)

$$\begin{aligned} f(v;\mu,\psi,\gamma) &= e^{-\psi(1-\gamma)(\frac{v}{\mu}-\frac{1}{\gamma})}\frac{1}{\pi}\sum_{k=1}^{\infty}(-1)^{k+1}\frac{[\psi(1-\gamma)]^{k(1-\gamma)}\mu^{k\gamma}\Gamma(k\gamma+1)}{\gamma^k k!} \\ &\quad \times v^{-k\gamma-1}\sin(k\gamma\pi), \end{aligned}$$

where $\mu > 0$, $\psi > 0$ and $0 < \gamma \leq 1$.

We use the restriction $\mathbb{E}[V] = \mu = 1$, such that $\mathbb{V}[V] = \frac{\mu^2}{\psi} = \frac{1}{\psi} := \theta$, where θ is interpretable as a measure of unobserved heterogeneity following the historical definition of frailty originally introduced by Vaupel et al. (VAUPEL; MANTON; STALLARD, 1979).

In order to eliminate the unobserved quantities, the random effect can be integrated out.

Thus, marginal survival function is given by

$$S(t;\mathbf{x}_1) = \mathbb{E}_V[S(t;\mathbf{x}_1,v)] = \int_0^\infty S(t;\mathbf{x}_1,v) f_v(v)dv = \mathscr{L}_v\left[-\log S_0(t;\mathbf{x}_1)\right],$$

where $f_v(\cdot)$ is the probability density of the corresponding frailty distribution, $S_0(\cdot)$ is the baseline survival function, and $\mathscr{L}_v[\cdot]$ denotes the Laplace transform of frailty distribution.

The unconditional survival and hazard functions in the PVF frailty model are expressed by, respectively,

$$S(t;\mathbf{x}_1) = \exp\left\{\frac{1-\gamma}{\gamma\theta}\left[1-\left(1+\frac{\lambda\theta}{1-\gamma}\int_0^t \exp\left\{-\exp(\alpha y+\mathbf{x}_1^\top\beta)\right\}dy\right)^\gamma\right]\right\} \tag{3.8}$$

and

$$h(t;\mathbf{x}_1) = \frac{\lambda\exp\left\{-\exp(\alpha t+\mathbf{x}_1^\top\beta)\right\}}{\left[1+\frac{\lambda\theta}{1-\gamma}\int_0^t \exp\left\{-\exp(\alpha y+\mathbf{x}_1^\top\beta)\right\}dy\right]^{1-\gamma}}. \tag{3.9}$$

Henceforth, we will refer to the model in which the survival function is shown in (3.8), as the extended CLL PVF frailty model, CLL PVF frailty model, or CLL PVF model. Note that the traditional CLL model (3.2) is obtained as $\theta \to 0$. If $\lambda = 1$ and $\theta \to 0$ the model (3.1) is derived. Besides, the CLL PVF model is flexible because it includes many other frailty models as special cases. For instance, when $\gamma \to 0$, the CLL gamma frailty model is obtained. The CLL inverse Gaussian is derived if $\gamma = 0.5$. The CLL positive stable frailty is a special case of the CLL PVF model in which some asymptotic considerations are necessary to show this fact. We refer the interested readers to Wienke (WIENKE, 2011).

The hazard function in (3.9) depends on the time; consequently, the CLL PVF model is also of non-proportional hazard. As does the CLL model, the CLL PVF model allows positive values for the time effect ($\alpha > 0$). Thus, the corresponding long-term survivors is

$$\begin{aligned} p(\mathbf{x}_1) &= \lim_{t\longrightarrow\infty} S(t;\mathbf{x}_1) \\ &= \lim_{t\longrightarrow\infty} \exp\left\{\frac{1-\gamma}{\gamma\theta}\left[1-\left(1+H_0(t;\mathbf{x}_1)\frac{\theta}{1-\gamma}\right)^\gamma\right]\right\} \in (0,1), \end{aligned} \tag{3.10}$$

where $H_0(\cdot;\mathbf{x}_1)$ is given by (3.3).

If parameter α is estimated to be positive, then the cure fractions for the CLL and CLL PVF frailty models can be obtained from (3.6) and (3.10), respectively. If parameter α is estimated to be negative, then there is no cure rate according to the two models, and functions (3.4) and (3.8) are proper survival functions.

A novel contribution of this work is also to incorporate explanatory variables in the extended CLL model (3.2) and CLL PVF frailty (3.7) models through parameter α, which is a more reasonable approach because it can directly reflect the effect of a treatment. As mentioned by Calsavara et al. (CALSAVARA *et al.*, 2019b) if a treatment effect is good in a specific group, then some patients will be cured, and the estimate for α will be $\alpha > 0$; otherwise, if the treatment is not sufficient, the estimate will be $\alpha < 0$, which to lead a non cured. Besides, it allows the intersection between the survival curves, which is commonly observed in clinical trials. Given this capacity, the extended CLL (3.2) and CLL PVF frailty (3.9) models are more flexible than standard approach (3.1) proposed by Milani, Diniz and Tomazella (2014).

As previously mentioned, explanatory variables are incorporated in the proposed models through of the effect time parameter α and in the traditional way in the hazard function with a set of two-covariate vectors, $\mathbf{x}_1 \in \mathbb{R}^p$ and $\mathbf{x}_2 \in \mathbb{R}^{q+1}$, such that $\mathbf{x}^\top = (\mathbf{x}_1^\top, \mathbf{x}_2^\top) \in \mathbb{R}^w$ is a w-dimensional covariate vector, where $w = p+q+1$. Importantly, parameter α can be estimated to be positive or negative. In this way to guarantee $\alpha \in \mathbb{R}$, we use an identity link function, such as

$$\alpha(\mathbf{x}_{2_i}) = \mathbf{x}_{2_i}^\top \alpha,$$

where $\mathbf{x}_{2_i}^\top = (1, x_{2_{i1}}, x_{2_{i2}}, \dots, x_{2_{iq}})$ and $\alpha^\top = (\alpha_0, \alpha_1, \dots, \alpha_q)$ are the sets of covariates and their regression coefficients, respectively. As suggest by Calsavara et al. (CALSAVARA *et al.*, 2020) if the researcher has prior knowledge about the variables that can be associated to long-term survivors they suggest link this subset variables to the α parameter.

An advantage of the proposed models, extended CLL and CLL PVF frailty models, over alternative models is the lack of assumption about long-term survivors' existence. Besides, the models allow that the time effect values lead to proper or improper distribution.

The CLL model (3.1) does not allow a flexible parametric fit for modeling phenomenon with non-monotone failure rates such as the unimodal and the bathtub-shaped failure rates commonly observed in biological studies and reliability. In this sense, the proposed model (3.9) has an advantage over the usual CLL model to accommodate monotone, and non-monotone failure rates can be applied in several problems in lifetime data analysis. The shape of the hazard and the survival functions obtained from CLL PVF frailty considering different parameter values are shown in Figure 5.

It is worth mentioning that the proposed model has intercept in the two components, α_0, and β_0, which leads to optimization problems, probably due to the non-identifiability of model parameters. In this sense, we will consider throughout the work intercept only in the component α.

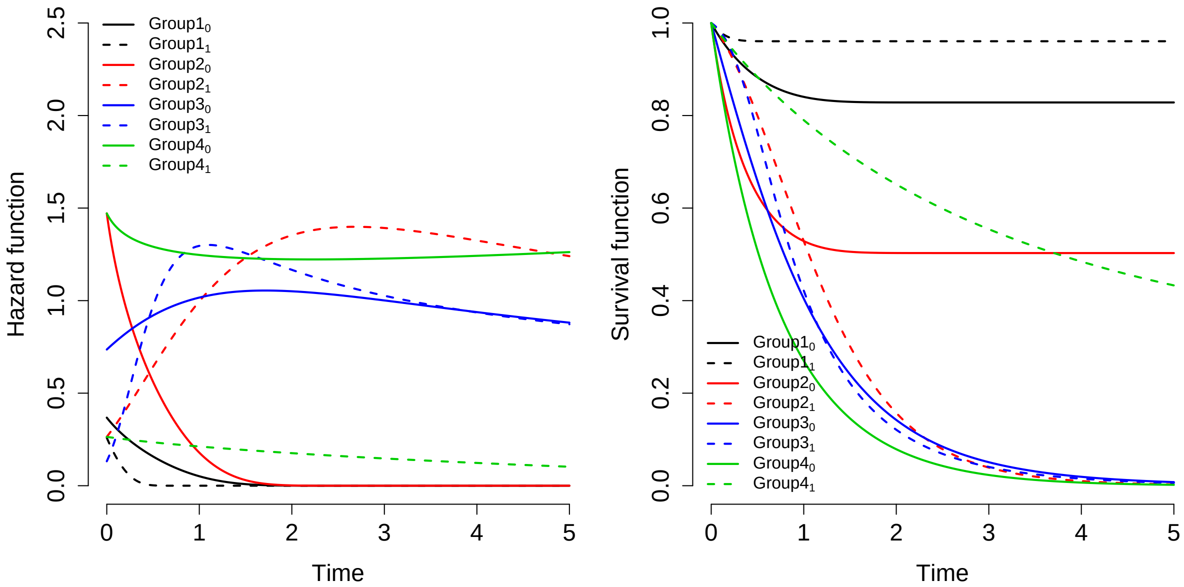

Figure 5 – Baseline hazard (left panel) and survival (right panel) functions from the CLL PVF frailty model. The parameter values used are: Group1, $\alpha_0 = 1$, $\alpha_1 = 2$, $\lambda = 1$, $\beta_0 = 0$, $\beta_1 = 0.3$, $\gamma = 0.7$, and $\theta = 2$; Group2, $\alpha_0 = 1$, $\alpha_1 = -2$, $\lambda = 4$, $\beta_0 = 0$, $\beta_1 = 1$, $\gamma = 0.7$, and $\theta = 1$; Group3, $\alpha_0 = -1$, $\alpha_1 = -2$, $\lambda = 2$, $\beta_0 = 0$, $\beta_1 = 1$, $\gamma = 0.7$, and $\theta = 0.5$; Group4, $\alpha_0 = -0.05$, $\alpha_1 = 0.1$, $\lambda = 4$, $\beta_0 = 0$, $\beta_1 = 1$, $\gamma = 0.9$, and $\theta = 0.5$; The subscript numerals indicate the values of the fixed covariates. (For interpretation of the references to color in this figure legend, the reader is referred to the online version of this article.)

3.4 Inference

Let us consider the situation when the time to event is not completely observed and it is subject to right censoring. Let $T > 0$ be a random variable representing the time until the occurrence of the event of interest. Let δ_i the censoring indicator variable, that is, $\delta_i = 0$ if the observed time is censored and $\delta_i = 1$ otherwise, $i = 1, \dots, n$. The observed dataset is $\mathbf{D} = (\mathbf{t}, \delta, \mathbf{X})$, where $\mathbf{t} = (t_1, \dots, t_n)^\top$ are the observed lifetimes, $\delta = (\delta_1, \dots, \delta_n)^\top$ are the censoring indicators, and $\mathbf{X}$ is a matrix containing the covariate information. Consider that T_is are independent and identically distributed random variables with survival and hazard functions specified, respectively, by $S(\cdot; \vartheta, \mathbf{x}_1, \mathbf{x}_2)$ and $h(\cdot; \vartheta, \mathbf{x}_1, \mathbf{x}_2)$, where ϑ denotes a vector of unknown parameters. We suppose that T is independent of the censoring time. Under non-informative censoring the likelihood function of ϑ is

$$L(\vartheta; \mathbf{D}) \propto \prod_{i=1}^{n} h(t_i; \vartheta, \mathbf{x}_{1i}, \mathbf{x}_{2i})^{\delta_i} S(t_i; \vartheta, \mathbf{x}_{1i}, \mathbf{x}_{2i}).$$

The corresponding log-likelihood function, $\ell(\vartheta) = \log L(\vartheta; \mathbf{D})$, is given by

$$\ell(\vartheta) \propto \sum_{i=1}^{n} \delta_i \log h(t_i; \vartheta, \mathbf{x}_{1i}, \mathbf{x}_{2i}) + \sum_{i=1}^{n} \log S(t_i; \vartheta, \mathbf{x}_{1i}, \mathbf{x}_{2i}).$$

Thus, for the extended CLL regression model the log-likelihood function for

$\vartheta = (\alpha, \lambda, \beta)^\top$ is

$$\begin{aligned}\ell(\vartheta) &= -\lambda \sum_{i=1}^{n} \int_0^t \exp\left\{-\exp\left(\mathbf{x}_{2i}^\top \alpha y + \mathbf{x}_{1i}^\top \beta\right)\right\} dy \\ &+ \log(\lambda) \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} \delta_i \exp\left(\mathbf{x}_{2i}^\top \alpha t_i + \mathbf{x}_{1i}^\top \beta\right),\end{aligned} \tag{3.11}$$

and for the CLL PVF frailty regression model the log-likelihood function for $\vartheta = (\alpha, \lambda, \beta, \theta, \gamma)^\top$ is

$$\begin{aligned}\ell(\vartheta) &= -(1-\gamma) \sum_{i=1}^{n} \delta_i \log\left[1 + \frac{\theta\lambda}{(1-\gamma)} \int_0^t \exp\left\{-\exp\left(\mathbf{x}_{2i}^\top \alpha y + \mathbf{x}_{1i}^\top \beta\right)\right\} dy\right] \\ &+ \sum_{i=1}^{n} \frac{1-\gamma}{\gamma\theta} \left(1 - \left[1 + \frac{\theta\lambda}{(1-\gamma)} \int_0^t \exp\left\{-\exp(\mathbf{x}_{2i}^\top \alpha y + \mathbf{x}_{1i}^\top \beta)\right\} dy\right]^{\gamma}\right) \\ &+ \log(\lambda) \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} \delta_i \exp\left(\mathbf{x}_{2i}^\top \alpha t_i + \mathbf{x}_{1i}^\top \beta\right).\end{aligned} \tag{3.12}$$

We numerically maximize the log-likelihood functions (3.11) and (3.12) to obtain the maximum likelihood estimates (MLEs) of parameters. There are many routines available for numerical maximization in the literature. We used the *optim* routine in the R software (R Core Team, 2020) for numerical maximization with the L-BFGS-B optimization method. In the numerical maximization, we also considered the *integrate* function to solve the integral (adaptive quadrature of functions of one variable) in the likelihood function once it has no analytical solution.

The asymptotic properties of maximum likelihood estimators are needed to obtain the confidence intervals and to test hypotheses about the model parameters. Under certain regularity conditions, $\widehat{\vartheta}$ has asymptotic multivariate normal distribution with mean ϑ and variance and covariance matrix $\Sigma(\widehat{\vartheta})$, which is estimated by

$$\widehat{\Sigma}(\widehat{\vartheta}) = \left\{ -\left.\frac{\partial^2 \ell(\vartheta; \mathbf{D})}{\partial \vartheta \partial \vartheta^\top}\right|_{\vartheta = \widehat{\vartheta}} \right\}.$$

Thus, an approximate $100(1-\alpha^*)\%$ confidence interval for ϑ_i is $\left(\widehat{\vartheta}_i - z_{\alpha^*/2}\sqrt{\widehat{\Sigma}^{ii}}, \widehat{\vartheta}_i + z_{\alpha^*/2}\sqrt{\widehat{\Sigma}^{ii}}\right)$, where $\widehat{\Sigma}^{ii}$ denotes the ith diagonal element of the inverse of $\widehat{\Sigma}$ and z_{α^*} denotes the $100(1-\alpha^*)$ percentile of the standard normal random variable.

The asymptotic normality assumption of MLEs holds only under certain regularity conditions, which are not easy to assess with our models. In the next section, we describe a simulation study performed to determine whether the usual asymptotic of the MLEs holds. Many authors have performed simulations to assess the asymptotic behavior of MLEs, especially when the analytical investigation is not trivial (MILANI; DINIZ; TOMAZELLA, 2014; CALSAVARA *et al.*, 2019a; CALSAVARA *et al.*, 2019b; CALSAVARA *et al.*, 2020).

3.5 Simulation study

In this section, we performed a simulation study in order to evaluate the performance of MLEs of the CLL PVF frailty (3.9) and extended CLL (3.2) models parameters considering different sample sizes. We also introduce two regression parameters in the effect time parameter α, that is, $\alpha(x) = \alpha_0 + \alpha_1 x$, where α_0 is the intercept and α_1 is the associated group variable. As previously mentioned, we did not incorporate the intercept in the other component due to the optimization problems. Thus, we will have just the regression coefficient β associated with the x variable. In addition, we considered an exponential distribution with the rate τ for the censoring times, in which τ is set to control the proportion of right-censored observations. Datasets (t_i, δ_i, x_i) from the CLL PVF frailty model and extended CLL model are generated using the following steps.

1. Determine the desired parameter values $\vartheta = (\alpha_0, \alpha_1, \lambda, \beta)^\top$ (extended CLL model) or $\vartheta = (\alpha_0, \alpha_1, \lambda, \beta, \theta, \gamma)^\top$ (CLL PVF frailty model);
2. For the ith subject, draw X_i ~Bernoulli(0.5) and $U_i^* \sim \text{Uniform}(0,1)$;
3. Determine the long-term survivors $p_i(x_i)$ according to the desired model;
4. Draw $C_i \sim \text{Exponential}\,(\tau)$, where τ is set to control the proportion of right-censored observations;
5. If $u_i^* < p_i(x_i)$, set $t_i^* = \infty$; otherwise, generate T_i^* from the CLL or CLL PVF frailty model, i.e., t_i^* as the root of $S(t_i^*;\vartheta) = 1 - u'$, where $U' \sim \text{Uniform}\,(0, 1 - p_i(x_i))$;
6. Let $t_i = \min\{t_i^*, c_i\}$;
7. If $t_i = t_i^*$, set $\delta_i = 1$; otherwise $\delta_i = 0$;
8. The dataset for the ith subject is $\{t_i, \delta_i, x_i\},\ i = 1, \ldots, n$.

We carried out an extensive Monte Carlo simulation considering sample sizes $n = 100, 200, 300, 500$ and 1000. For each combination of parameter values and sample size, we computed average MLEs of the parameters, their standard deviations (SDs), root mean square errors (RMSEs) of the MLEs of the parameters, and the empirical coverage probabilities (CPs) of 90% and 95% confidence intervals. The R software (R Core Team, 2020) was performed in all simulations with 1000 Monte Carlo runs. The L-BFGS-B algorithm of maximization was considered to estimates the parameters, which is an option of the *optim* function in R. In our simulation studies, we fixed the parameter $\gamma \to 0$ (CLL gamma frailty model) for all fitted models in order to corroborate with the results obtained in the application section.

Table 1 provides the results of the simulation studies of the CLL gamma frailty and extended CLL models. According to the results, the average estimates were close to fixed values as the sample size increased. Consequently, the bias gets to 0, regardless of the model parameters. The RMSEs and SDs decreased to 0 as the sample size increased. Besides, they are closer (RMSEs and SDs) when the sample size was $n \geq 300$. Empirical CPs for all parameters, regardless of the model, appeared to be reasonably close to the nominal level with increasing sample size. Considering the scenario of $n = 1000$, the empirical distributions of parameter estimates are shown in Figure 6. The plots indicate that the normal distribution provides reasonable approximations for estimator distributions.

Table 1 – Average of maximum likelihood estimates (AMLE), square roots of the mean squared errors (RMSEs), and standard deviations (SDs) of the maximum likelihood estimates, and empirical coverage probabilities (CPs) of 90% and 95% confidence intervals for the simulated data.

		CLL frailty model					Extended CLL model			
		α_0	α_1	λ	β	θ	α_0	α_1	λ	β
n		0.11	−0.09	2	1.1	1.5	0.18	−0.13	1.15	1.00
100	AMLE	0.127	-0.104	2.035	1.077	1.330	0.186	-0.131	1.181	0.988
	RMSE	0.045	0.043	0.731	0.149	0.685	0.033	0.040	0.238	0.142
	SD	0.042	0.041	0.730	0.148	0.664	0.033	0.040	0.236	0.141
	CP(90%)	0.918	0.926	0.872	0.907	0.892	0.922	0.906	0.907	0.903
	CP(95%)	0.971	0.969	0.912	0.948	0.944	0.964	0.961	0.948	0.942
200	AMLE	0.118	-0.096	2.022	1.088	1.422	0.183	-0.13	1.166	0.988
	RMSE	0.028	0.027	0.523	0.106	0.474	0.022	0.026	0.163	0.095
	SD	0.026	0.026	0.522	0.105	0.468	0.022	0.026	0.163	0.095
	CP(90%)	0.923	0.921	0.897	0.891	0.902	0.899	0.911	0.907	0.912
	CP(95%)	0.966	0.960	0.928	0.951	0.953	0.951	0.956	0.953	0.951
300	AMLE	0.116	-0.095	2.017	1.096	1.436	0.181	-0.129	1.159	0.999
	RMSE	0.023	0.023	0.435	0.088	0.408	0.018	0.021	0.135	0.080
	SD	0.022	0.022	0.435	0.088	0.403	0.018	0.021	0.135	0.080
	CP(90%)	0.893	0.892	0.879	0.877	0.891	0.904	0.892	0.895	0.892
	CP(95%)	0.940	0.943	0.921	0.942	0.934	0.958	0.945	0.944	0.944
500	AMLE	0.115	-0.094	1.991	1.093	1.442	0.181	-0.130	1.157	0.997
	RMSE	0.017	0.017	0.319	0.063	0.308	0.014	0.016	0.105	0.057
	SD	0.017	0.016	0.319	0.063	0.303	0.014	0.016	0.104	0.057
	CP(90%)	0.895	0.904	0.884	0.914	0.888	0.887	0.894	0.875	0.915
	CP(95%)	0.950	0.960	0.938	0.960	0.940	0.945	0.942	0.939	0.957
1000	AMLE	0.112	-0.092	2.004	1.097	1.480	0.180	-0.130	1.154	0.998
	RMSE	0.011	0.011	0.226	0.046	0.216	0.010	0.011	0.071	0.042
	SD	0.011	0.011	0.226	0.046	0.215	0.010	0.011	0.071	0.042
	CP(90%)	0.902	0.902	0.889	0.893	0.888	0.886	0.878	0.892	0.890
	CP(95%)	0.953	0.949	0.947	0.950	0.941	0.940	0.939	0.950	0.950

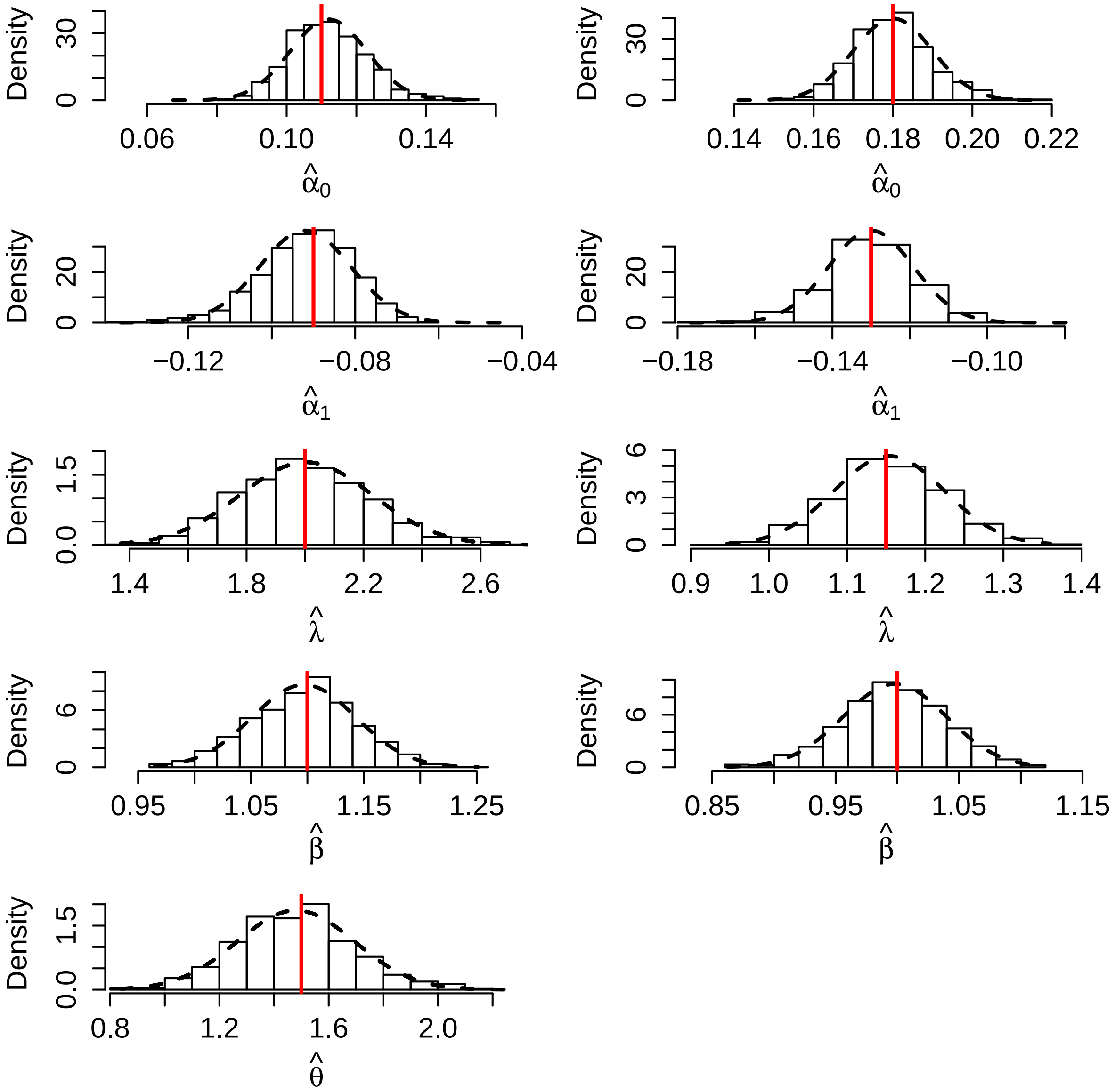

Figure 6 – Histogram of MLEs of parameters for sample size $n = 1000$ and the fixed parameter value (red line). Left panel: CLL gamma frailty model. Right panel: Extended CLL model.

CHAPTER 4

APPLICATION

In this chapter, we consider a real cancer dataset to illustrate the applicability of the proposed models. We fitted the extended CLL and CLL PVF frailty models and their particular models to the melanoma cancer dataset and compared them with survival curve estimates obtained using the Kaplan-Meier estimator. We provide the MLEs, standard error, 95% confidence interval estimates for the parameters, and AIC criterion value. Estimates of the 95% percentile bootstrap confidence interval for the long-term survivors' parameters were obtained using the nonparametric bootstrap technique with 100 bootstrap samples.

4.1 Melanoma cancer data

The melanoma data are part of a study about skin cancer where the 6752 patients diagnosed with melanoma in the state of São Paulo, Brazil, were included in the study between 2000 and 2014, with follow-up conducted until 2018. The event of interest was defined as death attributed to cancer; a total of 1914 (28.3%) events occurred during the follow-up period. The overall melanoma-specific survival is shown in Figure 1. The goal was to assess the impact of observed covariates on specific survival and the long-term survivors using the proposed models. The estimated survival functions for each observed covariate are shown in Figure 2.

Melanoma is one of the best known by the population, but skin carcinomas are more incidents than melanoma. Worldwide, the staging system proposed by the American Joint Committee on Cancer (AJCC) is commonly used for melanoma and other solid tumors. The early clinical stages (I or II) have been associated with a better prognosis once it corresponds to the melanoma limited to the skin, and these patients are typically treated with surgery. In the early stages, the vast majority will be alive after ten years of follow-up. Patients in the clinical stages III the surgery is associated with radiotherapy, or some modality of systemic treatment such as immunotherapy or targeted therapies (EGGERMONT; DUMMER, 2017). Clinical stage IV has the worst prognosis once it corresponds to metastatic disease. Some of these patients may undergo surgery at some time, but they will more likely need systemic treatment (PUZA *et al.*, 2019). In our study, 4313 (72.1%) of the patients who underwent surgery were in stage clinical I or II, while 529 (68.6%) patients who did not undergo surgery were in stage clinical III or IV. As there is an association between the surgery and disease stage (clinical stage), we will take into account only the treatment variable as the covariate and age at diagnosis and gender variables.

This dataset was initially studied by Calsavara *et al.* (2020), where they evaluated only the effect of surgery on the hazard function using a non-proportional hazards model with a frailty term. Recently, Molina *et al.* (2021) and Rodrigues *et al.* (2021) also considered the same data set, but in both papers, the main goal was to assess the effect of the clinical stage on the melanoma-specific survival instead of the surgery variable.

As mentioned in Chapter 2, the proportionality assumption is questionable for surgery covariate according to the plot of log cumulative baseline hazard rates against time (follow-up period) as shown in Figure 2. In Figure **??** we provide in Figure 7 a plot of standardized Schoenfeld residuals against time for these covariates obtained from the fitted Cox regression model. The results of proportional hazards assumption testing for the fitted Cox regression model (GRAMBSCH; THERNEAU, 1994) are displayed in Table 2; they provided strong evidence that the surgery variable had a non-constant effect over time, while the age at diagnosis and gender there is evidence of effect constant over time.

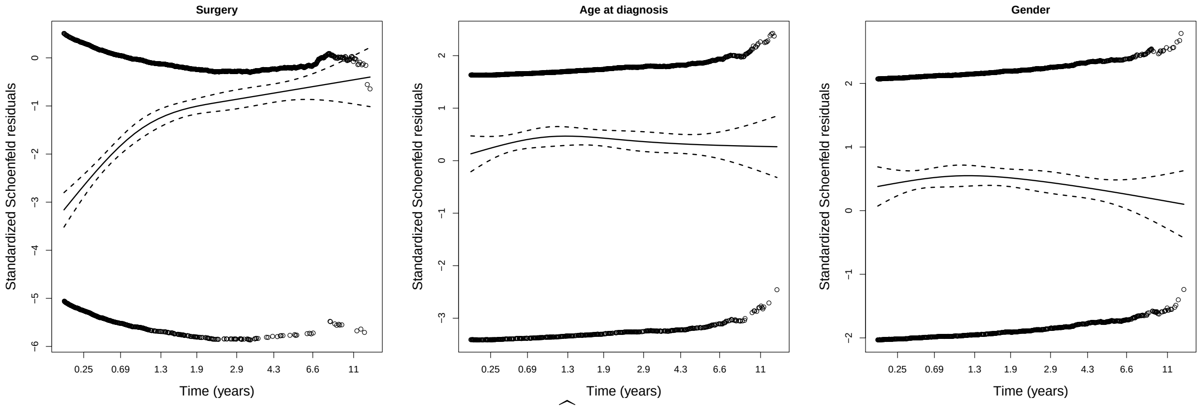

Figure 7 – Standardized Schoenfeld residuals+$\widehat{\beta}$ for the covariate surgery (left panel), age at diagnosis (middle panel) and gender (right panel) plotted from fitted Cox model.

Table 2 – Test of proportional hazards assumption.

Variable	ρ	χ^2	p-value
Surgery	0.287	150	<0.0001
Age at diagnosis	0.002	0.008	0.926
Gender	-0.030	1.680	0.195

To evaluate the observed covariates' effect in the hazard function and the effect time, we fitted the extended CLL and CLL PVF frailty models to the dataset. For illustrative purposes, we link parameter α to covariates through an identity link function. Thus,

$$\alpha(x_i) = \alpha_0 + x_i\alpha_1,$$

where x_i indicates the covariate associated to the patient for $i = 1,\ldots,6752$; and $\alpha^\top = (\alpha_0, \alpha_1)$ represents the regression coefficients. The results of the fitted extended CLL and CLL PVF frailty models are given in Table 3. Notice that the estimate of γ is close to zero indicating that a CLL gamma frailty model can be considered. In this sense, we also fitted to the dataset the main special cases, CLL inverse Gaussian ($\gamma = 0.5$) and gamma ($\gamma \to 0$) frailty models, and the results are given in Table 4. According to the AIC value, the CLL gamma frailty model seems to be the better choice among the four models.

The results suggest a significant surgery effect in the lifetime regardless of the model, as the 95% confidence interval the β does not include 0. Besides, the time effect measure differs between groups (α_0 and α_1 are significant), except age at diagnosis. Note that $\widehat{\alpha}_0 > 0$ and $\widehat{\alpha}_0 + \widehat{\alpha}_1 > 0$ in the four models, which means that the distributions were improper, leading to long-term survivors in the three observed covariates.

As mentioned previously, of the four fitted models, the CLL gamma frailty model gave the best fit according to the AIC value. However, the CLL PVF frailty model can also be considered once the difference between AIC values is slight and the parameter estimates are similar.

Considering the AIC criterion, max $\ell(\cdot)$ values, and the number of parameters in the model, we select the CLL gamma frailty as our working model. Accordingly, we focused exclusively on an interpretation of CLL gamma frailty model parameters. Note that $\widehat{\theta} = 1.492$, which indicates a reasonable degree of unobserved heterogeneity in the sample when surgery is considered in the model. The amount of estimated unobserved heterogeneity when the age at diagnostic and gender are considered independently in the model is $\widehat{\theta} = 2.213$ and $\widehat{\theta} = 2.111$, respectively.

In addition, the estimated time effects were $\widehat{\alpha}_0 = 0.114$; CI(95%) $= [0.083; 0.145]$ in the no surgery group and $\widehat{\alpha}_0 + \widehat{\alpha}_1 = 0.029$; CI(95%) $= [0.022; 0.038]$ in the surgery group. These estimates evidence that the time effect is not the same in both groups. As the time effects are positive, the model suggests that there are long-term survivors, as can be seen in the estimated proportions, $\widehat{p}_0 = 0.278$; bootstrap CI(95%) $= [0.236; 0.323]$ (no surgery) and $\widehat{p}_1 = 0.616$; bootstrap CI(95%) $= [0.560; 0.629]$ (surgery).

For the age at diagnosis, the estimated time effects $\widehat{\alpha}_0 = 0.084$; CI(95%) $= [0.064; 0.104]$ in the younger patients and $\widehat{\alpha}_0 + \widehat{\alpha}_1 = 0.102$; CI(95%) $= [0.070; 0.133]$ in the older patients. As the time effect are both positive, the model suggests that there are long-term survivors; the estimated proportions are, $\widehat{p}_0 = 0.662$; bootstrap CI(95%) $= [0.629; 0.686]$ (younger patients) and $\widehat{p}_1 = 0.556$; bootstrap CI(95%) $= [0.509; 0.575]$ (older patients). The estimated time effects for the gender were: $\widehat{\alpha}_0 = 0.078$; CI(95%) $= [0.061; 0.096]$ in the female patients and $\widehat{\alpha}_0 +$

$\widehat{\alpha}_1 = 0.126$; CI(95%) = [0.090;0.164] in the male patients. As the time effects are positive the estimated proportions are $\widehat{p}_0 = 0.650$; bootstrap CI(95%) = [0.607;0.679] (female patients) and $\widehat{p}_1 = 0.532$; bootstrap CI(95%) = [0.496;0.570] (male patients).

Overall, the models reasonably fit Kaplan-Meier curves. However, the CLL frailty model enables quantifying unobserved heterogeneity, which is of great importance in clinical practice, once those significant covariates were not observed, such as Breslow thickness, ulceration, and Mitotic rate.

Table 3 – Maximum likelihood estimate (MLEs), standard error (SE), 95% asymptotic confidence intervals (CI), AIC value obtained for the extended CLL and CLL PVF frailty models categorized by surgery, age and gender fitted for the melanoma dataset.

Model	Extended CLL				CLL PVF frailty			
			CI 95%				CI 95%	
Parameter	MLE	SE	Lower	Upper	MLE	SE	Lower	Upper
α_0	0.182	0.008	0.166	0.199	0.125	0.015	0.095	0.154
$\alpha_{1\,(\text{Yes})}$	-0.131	0.009	-0.149	-0.113	-0.094	0.014	-0.122	-0.066
λ	1.136	0.061	1.017	1.255	2.064	0.202	1.669	2.460
$\beta_{(\text{Yes})}$	0.969	0.024	0.922	1.016	1.143	0.032	1.080	1.206
γ	-	-	-	-	0.130	0.083	0.0001	0.294
θ	-	-	-	-	1.547	0.222	1.112	1.983
$2\max\ell(\cdot)$	$-13,404.36$				$-13,330.72$			
AIC	13,412.36				13,342.72			
α_0	0.118	0.007	0.105	0.131	0.084	0.010	0.064	0.104
$\alpha_{1\,(\text{Older})}$	0.024	0.012	0.001	0.048	0.018	0.018	-0.017	0.053
λ	0.221	0.011	0.200	0.243	0.259	0.017	0.225	0.294
$\beta_{(\text{Older})}$	-0.331	0.070	-0.468	-0.193	-0.550	0.127	-0.798	-0.302
γ	-	-	-	-	0.001	0.014	0.0001	0.028
θ	-	-	-	-	2.201	0.310	1.593	2.809
$2\max\ell(\cdot)$	$-13,950.93$				$-13,892.63$			
AIC	13,958.93				13,904.63			
α_0	0.112	0.006	0.101	0.123	0.078	0.009	0.060	0.096
$\alpha_{1\,(\text{Male})}$	0.050	0.012	0.025	0.074	0.050	0.021	0.010	0.090
λ	0.216	0.009	0.199	0.234	0.253	0.014	0.226	0.280
$\beta_{(\text{Male})}$	-0.518	0.074	-0.664	-0.372	-0.869	0.158	-1.179	-0.560
γ	-	-	-	-	0.002	0.025	0.0001	0.051
θ	-	-	-	-	2.108	0.293	1.534	2.682
$2\max\ell(\cdot)$	$-13,904.51$				$-13,846.07$			
AIC	13,912.51				13,858.07			

Figure 8 shows the estimated survival and hazard functions from the extended CLL and CLL gamma frailty models for each observed covariate. In both models, but more so in the CLL gamma frailty model, the survival function estimates are close to the Kaplan-Meier curves. Besides, the hazard function curves are higher for patients who did not undergo surgery,

Table 4 – Maximum likelihood estimate (MLEs), standard error (SE), 95% asymptotic confidence intervals (CI), AIC value obtained for the CLL Gamma and CLL IG frailty models categorized by surgery, age and gender fitted for the melanoma dataset.

Model	CLL Gamma frailty model				CLL IG frailty model			
			CI 95%				CI 95%	
Parameter	MLE	SE	Lower	Upper	MLE	SE	Lower	Upper
α_0	0.114	0.016	0.083	0.145	0.146	0.011	0.124	0.167
$\alpha_{1\,(\mathrm{Yes})}$	-0.085	0.015	-0.114	-0.055	-0.112	0.011	-0.133	-0.090
λ	2.013	0.184	1.652	2.373	1.769	0.158	1.458	2.079
$\beta_{(\mathrm{Yes})}$	1.139	0.031	1.079	1.199	1.094	0.030	1.036	1.152
γ	-	-	-	-	-	-	-	-
θ	1.492	0.197	1.107	1.878	1.565	0.318	0.942	2.188
$2\max \ell(\cdot)$	$-13,326.22$				$-13,348.52$			
AIC	$13,336.22$				$13,358.52$			
α_0	0.084	0.010	0.064	0.104	0.091	0.009	0.073	0.109
$\alpha_{1\,(\mathrm{Older})}$	0.018	0.018	-0.018	0.053	0.027	0.016	-0.005	0.058
λ	0.260	0.018	0.225	0.294	0.265	0.018	0.229	0.300
$\beta_{(\mathrm{Older})}$	-0.551	0.127	-0.800	-0.302	-0.512	0.114	-0.735	-0.288
γ	-	-	-	-	-	-	-	-
θ	2.213	0.311	1.604	2.821	2.662	0.518	1.646	3.678
$2\max \ell(\cdot)$	$-13,892.63$				$-13,899.32$			
AIC	$13,902.63$				$13,909.32$			
α_0	0.078	0.009	0.061	0.096	0.085	0.008	0.069	0.101
$\alpha_{1\,(\mathrm{Male})}$	0.048	0.020	0.008	0.088	0.057	0.018	0.022	0.092
λ	0.254	0.014	0.227	0.281	0.260	0.015	0.230	0.289
$\beta_{(\mathrm{Male})}$	-0.857	0.155	-1.161	-0.553	-0.783	0.132	-1.041	-0.524
γ	-	-	-	-	-	-	-	-
θ	2.111	0.293	1.538	2.685	2.575	0.504	1.587	3.564
$2\max \ell(\cdot)$	$-13,846.04$				$-13,853.00$			
AIC	$13,856.04$				$13,863.00$			

mainly in the first five years of follow-up, regardless of models. In both models, the fitted hazard functions decrease over time; the curves also cross over time. Such crossing can not occur in the traditional CLL model (considering the effect time α equals in both groups) proposed by Milani et al. (MILANI; DINIZ; TOMAZELLA, 2014) disadvantage. The inclusion of a covariate in the α parameter allowed the quantification of each group of patients' effect and allowed the curves to cross, as can be seen in the estimated hazard function from the models in Figure 9.

We also fitted a full model considering all risk factors previously mentioned. The results of the fitted proposed models are given in Table 5. According to the AIC criterion values, frailty models seem to be the better choice. The smallest value occurred for the CLL gamma frailty model. Among the observed covariates considered in the models, there is evidence that surgery, age at diagnosis, and gender are important factors to explain the failure rate, as the 95%

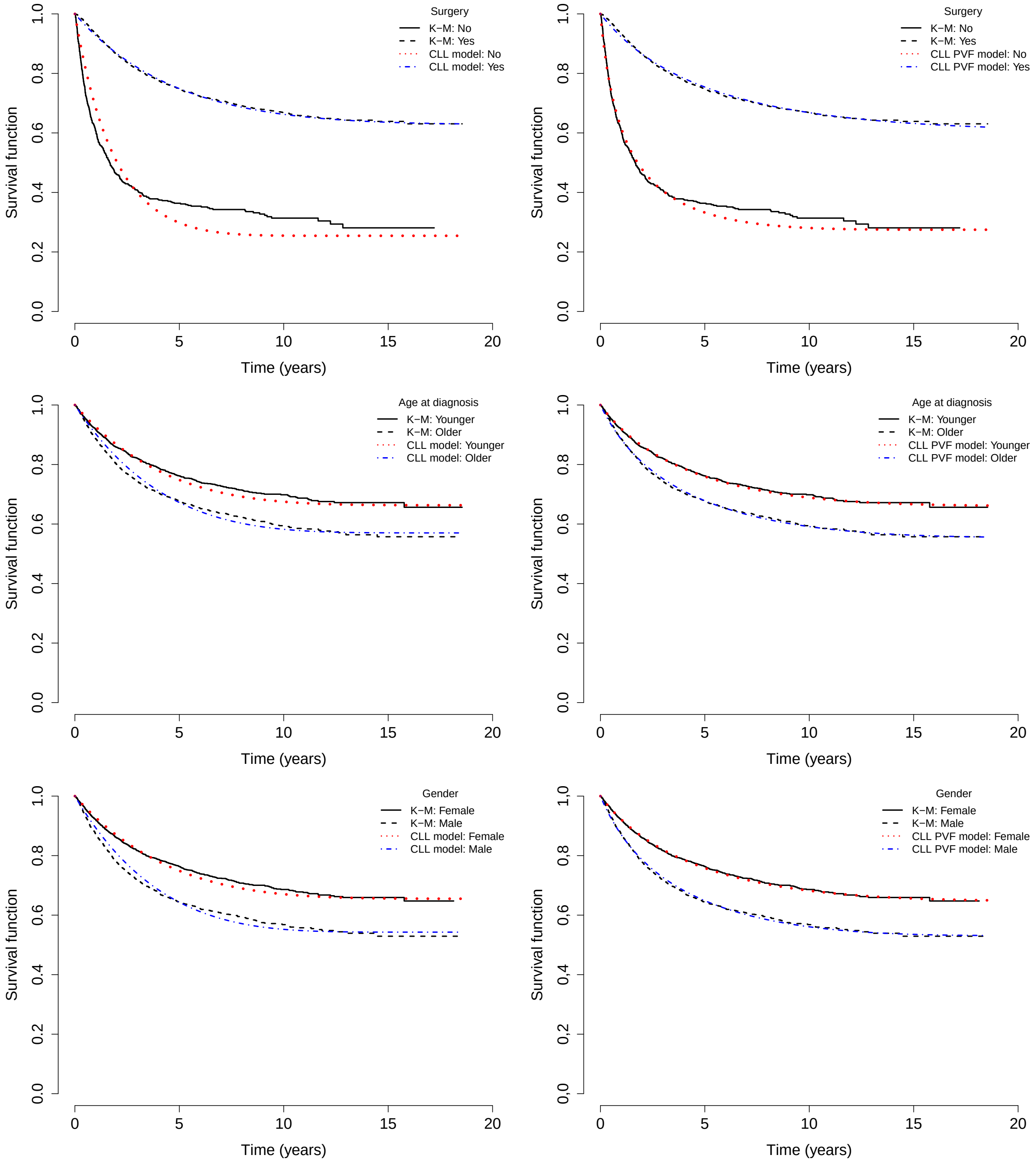

Figure 8 – Estimated survival curve obtained via Kaplan-Meier (black line) for melanoma dataset, and estimated survival function according to extended CLL model (left panel) and CLL PVF frailty model (right panel) for surgery (top panel), age at diagnosis (middle panel) and gender (bottom panel).

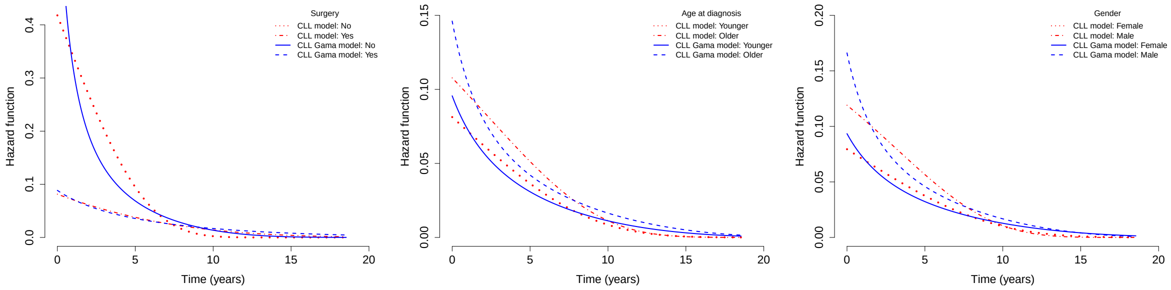

Figure 9 – Estimated hazard function according to extended CLL model and CLL gamma frailty model for surgery (left panel), age at diagnosis (middle panel) and gender (right panel).

confidence interval of the coefficients $\beta^\top = (\beta_1, \beta_2, \beta_3)$ do not include 0, regardless of the fitted models. There is a significant effect time only for the surgery covariate regarding frailty models, which indicates that the effect time is different between groups (α_0 and α_1 are significant). Considering the AIC criterion, max $\ell(\cdot)$ values, and the number of parameters in the model, we select the CLL gamma frailty model as our working model. Note that $\widehat{\theta} = 1.457$, which indicates a reasonable degree of unobserved heterogeneity in the sample. In addition, the effect time in the no surgery group was $\widehat{\alpha}_0 = 0.125$; CI(95%) $= [0.093; 0.156]$ and $\widehat{\alpha}_0 + \widehat{\alpha}_1 = 0.027$; CI(95%) $= [0.015; 0.038]$ in the surgery group. As the time effects are positive the model suggests that there are long-term survivors.

The proposed frailty model allows quantifying the amount of unobserved heterogeneity, which is of great importance in clinical practice. We tested the suitability of the frailty term in the CLL frailty model using the likelihood ratio test given by, $\Lambda = 2\{\ell(\widehat{\vartheta}) - \ell(\widehat{\vartheta}_0)\}$, where $\widehat{\vartheta}_0$ is the maximum likelihood estimator of $\widehat{\vartheta}$ under the null hypothesis H, where $H : \theta = 0$. Maller and Zhou (MALLER; ZHOU, 1996) showed that the statistical distribution Λ is a mixture in proportions $50\%/50\%$ of a chi-squared distribution with one degree of freedom and a point mass at 0, that is $\mathbb{P}[\Lambda \leq \xi] = 0.5 + 0.5\mathbb{P}[\chi^2_1 \leq \xi]$ under certain regularity conditions. We obtained $\Lambda = 58.45$ (p-value< 0.0001), which provides evidence in favor of the inclusion of the frailty term.

As mentioned previously, the CLL gamma frailty model gave the best fit according to the AIC criterion value. However, the CLL PVF frailty model can also be considered once the difference between AIC values is slight and the parameter estimates are similar.

The inclusion of the scalar λ in the traditional model (3.1), as well as in the extended CLL frailty model (3.9) improved the flexibility of the model, as can be seen in the estimate $\widehat{\lambda} \neq 1$.

Figure 10 shows the estimated survival function according to the CLL gamma frailty model for all combinations of covariates surgery, gender, and age at diagnosis. Table 6 shows the estimated long-term survivors for all combinations of observed covariates. In general, patients who have undergone surgery have better survival than those who did not undergo surgery, as expected, since most patients who underwent this treatment had an early diagnosis. In addition,

Table 5 – Maximum likelihood estimate (MLEs), standard error (SE), 95% asymptotic confidence intervals (CI), AIC value obtained for the CLL PVF frailty model and their special cases categorized by surgery, age and gender fitted for the melanoma dataset.

Model	Extended CLL				CLL PVF frailty			
			CI 95%				CI 95%	
Parameter	MLE	SE	Lower	Upper	MLE	SE	Lower	Upper
α_0	0.184	0.010	0.165	0.203	0.135	0.015	0.105	0.165
$\alpha_{1(\text{Yes})}$	-0.145	0.010	-0.164	-0.125	-0.110	0.015	-0.139	-0.081
$\alpha_{2(\text{Older})}$	0.007	0.007	-0.006	0.020	0.005	0.007	-0.009	0.018
$\alpha_{3(\text{Male})}$	0.019	0.007	0.006	0.032	0.007	0.007	-0.007	0.021
λ	0.916	0.057	0.805	1.027	1.347	0.128	1.095	1.598
$\beta_{1(\text{Yes})}$	1.093	0.028	1.037	1.148	1.217	0.035	1.148	1.286
$\beta_{2(\text{Age})}$	-0.143	0.031	-0.204	-0.083	-0.163	0.031	-0.225	-0.102
$\beta_{3(\text{Male})}$	-0.213	0.029	-0.270	-0.156	-0.203	0.030	-0.262	-0.143
γ	-	-	-	-	0.329	0.126	0.082	0.576
θ	-	-	-	-	1.511	0.266	0.989	2.033
$2\max \ell(\cdot)$	$-13,279.22$				$-13,218.77$			
AIC	$13,295.22$				$13,238.77$			

Model	CLL Gamma frailty				CLL IG frailty			
α_0	0.125	0.016	0.093	0.156	0.153	0.012	0.130	0.176
$\alpha_{1(\text{Yes})}$	-0.098	0.016	-0.129	-0.068	-0.125	0.012	-0.148	-0.102
$\alpha_{2(\text{Older})}$	0.001	0.007	-0.012	0.014	0.003	0.007	-0.010	0.016
$\alpha_{3(\text{Male})}$	0.003	0.007	-0.011	0.016	0.009	0.007	-0.004	0.022
λ	1.655	0.163	1.335	1.975	1.413	0.134	1.150	1.676
$\beta_{1(\text{Yes})}$	1.267	0.034	1.201	1.334	1.221	0.034	1.156	1.287
$\beta_{2(\text{Older})}$	-0.145	0.030	-0.203	-0.087	-0.149	0.031	-0.209	-0.089
$\beta_{3(\text{Male})}$	-0.177	0.028	-0.233	-0.121	-0.194	0.029	-0.251	-0.137
γ	-	-	-	-	-	-	-	-
θ	1.457	0.194	1.078	1.837	1.468	0.302	0.877	2.059
$2\max \ell(\cdot)$	$-13,204.43$				$-13,225.61$			
AIC	$13,222.43$				$13,243.61$			

younger female patients who have undergone surgery have better survival than those who did not undergo surgery, regardless of age at diagnosis. Female patients exhibited slightly better long-term survivors than male patients with the same treatment (surgery). Meanwhile, in the absence of surgery, the estimated long-term survivors are worse, mainly in older male patients. These results are in line with those observed in Calsavara *et al.* (2020), Molina *et al.* (2021), Rodrigues *et al.* (2021).

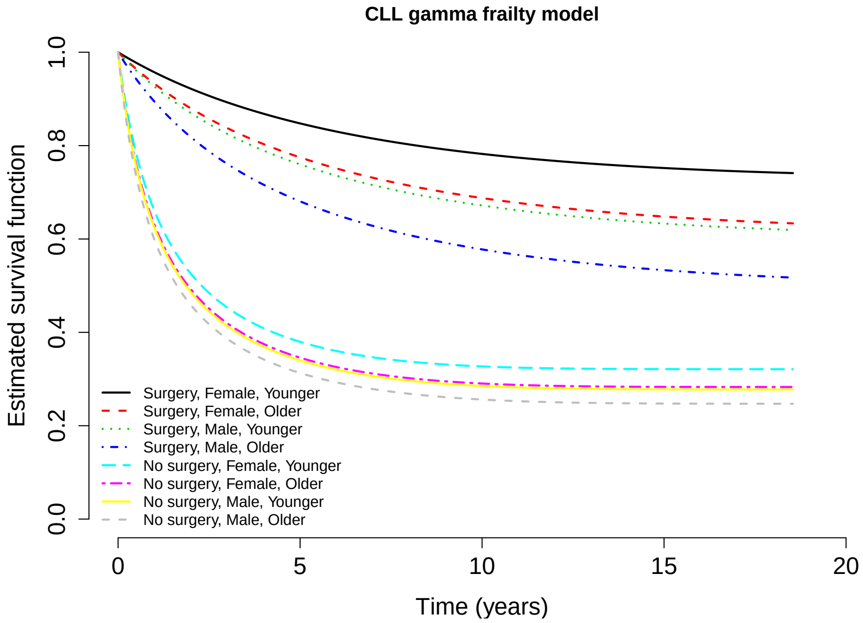

Figure 10 – Estimated survival function according to CLL gamma frailty model for all combinations of the covariates surgery, gender and age at diagnosis.

Table 6 – Estimated long-term survivors according to the CLL gamma frailty model stratified by surgery, gender and age at diagnosis.

	Surgery				No surgery			
Long-term	Female		Male		Female		Male	
survivors	Younger	Older	Younger	Older	Younger	Older	Younger	Older
p	0.741	0.633	0.619	0.517	0.321	0.283	0.278	0.247

The results finding in our study are consistent with those observed in routine clinical practice. Surgery, gender, and age at diagnosis have already been reported as prognostic factors, suggesting that younger patients and women have a better prognosis (SABEL *et al.*, 2005; BALCH *et al.*, 2014; BALCH *et al.*, 2001; BALCH *et al.*, 2009; GERSHENWALD *et al.*, 2017). As previously mentioned, those patients who had early diagnosis were in great majority treated with surgery, which is associated with a better prognosis. The estimated curves shown in this dissertation are very similar to those presented in the three latest updates of the AJCC staging system for melanoma (BALCH *et al.*, 2001; BALCH *et al.*, 2009; GERSHENWALD *et al.*, 2017).

CHAPTER

5

FINAL REMARKS

5.1 Conclusion

In this dissertation, we extended the complementary log-log hazard model (CLL) proposed by Milani et al. (MILANI; DINIZ; TOMAZELLA, 2014) with the inclusion of a scalar parameter λ in the hazard function allowing the new hazard function is not limited in the unit interval. We also proposed a generalized of the extended CLL model with a PVF frailty term for right-censored data. An advantage of the proposed model over alternatives is that it does not make assumptions about the existence of the long-term survivors, once the parameter α value has led to proper $(\alpha \leq 0)$ or improper $(\alpha > 0)$ distribution; this makes the model flexible and applicable to situations presence and absence long-term survivors. If parameter α is estimated to be positive, then the long-term survivors are computed as a function of the CLL model parameters. Besides, the inclusion of a frailty term in the hazard function quantifies unobserved heterogeneity employing the parameter θ. In our simulation study, conducted to illustrate the various properties of the MLEs of the parameters, the bias and RMSEs appeared to trend reasonably close to 0 as the sample size increased. The simulation study showed that the CLL frailty model is not indicated for small $(n \leq 100)$ samples. The practical relevance and applicability of the proposed models were demonstrated using a real melanoma dataset, where surgery, age, and gender covariates were essential factors to explain the failure rate and the time effect, which was different only for the surgery. Although further research on this approach must be conducted, our initial results suggest that this model enhances the analysis of non-proportional hazards in the presence or absence of long-term survivors.

5.2 Future work

The present study leaves some open topics to be addressed in the future. For instance, we may consider developing more simulation studies considering several values of unobserved heterogeneity in the sample, incorporating two or more covariates in the two components, study the impact of the MLEs when there are not immunes in the population, or when a subgroup has long-term survivors and the others not, and developing a power analysis for planing sample size.

BIBLIOGRAPHY

AALEN, O. O. Heterogeneity in survival analysis. **Statistics in Medicine**, v. 7, p. 1121–1137, 1988. Citation on page 20.

ANDERSEN, P. K.; BORGAN, O.; GILL, R. D.; KEIDING, N. **Statistical models based on counting processes**. [S.l.]: Springer Science & Business Media, 2012. Citation on page 20.

ANDRADE, C. T. d.; MAGEDANZ, A. M. P. C. B.; ESCOBOSA, D. M.; TOMAZ, W. M.; SANTINHO, C. S.; LOPES, T. O.; LOMBARDO, V. The importance of a database in the management of healthcare services. **Einstein (São Paulo)**, v. 10, p. 360–365, 2012. Citation on page 21.

BALCH, C. M.; BUZAID, A. C.; SOONG, S.-J.; ATKINS, M. B.; CASCINELLI, N.; COIT, D. G.; FLEMING, I. D.; GERSHENWALD, J. E.; JR, A. H.; KIRKWOOD, J. M. *et al.* Final version of the american joint committee on cancer staging system for cutaneous melanoma. **Journal of Clinical Oncology**, v. 19, p. 3635–3648, 2001. Citation on page 51.

BALCH, C. M.; GERSHENWALD, J. E.; SOONG, S.-j.; THOMPSON, J. F.; ATKINS, M. B.; BYRD, D. R.; BUZAID, A. C.; COCHRAN, A. J.; COIT, D. G.; DING, S. *et al.* Final version of 2009 AJCC melanoma staging and classification. **Journal of Clinical Oncology**, American Society of Clinical Oncology, v. 27, p. 6199, 2009. Citation on page 51.

BALCH, C. M.; THOMPSON, J. F.; GERSHENWALD, J. E.; SOONG, S.-j.; DING, S.; MCMASTERS, K. M.; COIT, D. G.; EGGERMONT, A. M.; GIMOTTY, P. A.; JOHNSON, T. M. *et al.* Age as a predictor of sentinel node metastasis among patients with localized melanoma: an inverse correlation of melanoma mortality and incidence of sentinel node metastasis among young and old patients. **Annals of Surgical Oncology**, v. 21, p. 1075–1081, 2014. Citation on page 51.

BALKA, J.; DESMOND, A. F.; MCNICHOLAS, P. D. Review and implementation of cure models based on first hitting times for Wiener processes. **Lifetime Data Analysis**, v. 15, p. 147–176, 2009. Citation on page 21.

____. Bayesian and likelihood inference for cure rates based on defective inverse Gaussian regression models. **Journal of Applied Statistics**, v. 38, p. 127–144, 2011. Citation on page 21.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, v. 47, n. 259, p. 501–515, 1952. Citation on page 20.

BERTOLLI, E.; FRANKE, V.; CALSAVARA, V. F.; MACEDO, M. P. de; PINTO, C. A. L.; HOUDT, W. J. van; WOUTERS, M. W.; NETO, J. P. D.; AKKOOI, A. C. van. Validation of a nomogram for non-sentinel node positivity in melanoma patients, and its clinical implications: a brazilian–dutch study. **Annals of Surgical Oncology**, v. 26, n. 2, p. 395–405, 2019. Citation on page 24.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, v. 11, n. 1, p. 15–53, 1949. Citation on page 20.

CALSAVARA, V. F.; MILANI, E. A.; BERTOLLI, E.; TOMAZELLA, V. Long-term frailty modeling using a non-proportional hazards model: Application with a melanoma dataset. **Statistical Methods in Medical Research**, v. 29, p. 2100–2118, 2020. Citations on pages 19, 20, 21, 24, 28, 37, 39, 44, and 50.

CALSAVARA, V. F.; RODRIGUES, A. S.; ROCHA, R.; TOMAZELLA, V.; LOUZADA, F. Defective regression models for cure rate modeling with interval-censored data. **Biometrical Journal**, v. 61, p. 841–859, 2019. Citations on pages 21 and 39.

CALSAVARA, V. F.; RODRIGUES, A. S.; ROCHA, R.; LOUZADA, F.; TOMAZELLA, V.; SOUZA, A. C.; COSTA, R. A.; FRANCISCO, R. P. Zero-adjusted defective regression models for modeling lifetime data. **Journal of Applied Statistics**, v. 46, p. 2434–2459, 2019. Citations on pages 21, 36, and 39.

CALSAVARA, V. F.; RODRIGUES, A. S.; TOMAZELLA, V. L. D.; CASTRO, M. de. Frailty models power variance function with cure fraction and latent risk factors negative binomial. **Communications in Statistics-Theory and Methods**, v. 46, p. 9763–9776, 2017. Citation on page 20.

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. Citation on page 20.

CLAYTON, D. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. **Biometrika**, JSTOR, v. 65, p. 141–151, 1978. Citation on page 20.

Coordenação de Prevenção e Vigilância. **Instituto Nacional de Câncer José Alencar Gomes da Silva. Estimativa 2018: Incidência de Câncer no Brasil. Coordenação de Prevenção e Vigilância – Rio de Janeiro**. 2017. <http://www1.inca.gov.br/estimativa/2018/>. Citation on page 19.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society B**, v. 34, p. 187–220, 1972. Citations on pages 19, 27, and 34.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap methods and their application**. [S.l.]: Cambridge University Press, 1997. Citation on page 32.

EGGERMONT, A. M.; DUMMER, R. The 2017 complete overhaul of adjuvant therapies for high-risk melanoma and its consequences for staging and management of melanoma patients. **European Journal of Cancer**, v. 86, p. 101–105, 2017. Citation on page 43.

ERVIK, M.; LAM, F.; FERLAY, J.; MERY, L.; SOERJOMATARAM, I.; BRAY, F. *et al.* Cancer Today Lyon, France: International Agency for Research on Cancer. 2016. **Cancer Today. Available from: http://gco.iarc.fr/today, accessed [10/09/2020]**, 2020. Citation on page 19.

ETEZADI-AMOLI, J.; CIAMPI, A. Extended hazard regression for censored survival data with covariates: a spline approximation for the baseline hazard function. **Biometrics**, JSTOR, v. 43, p. 181–192, 1987. Citation on page 20.

FONSECA, I. B.; LINDOTE, M. V. N.; MONTEIRO, M. R.; FILHO, E. D.; PINTO, C. A. L.; JAFELICCI, A. S.; LÔBO, M. de M.; CALSAVARA, V. F.; BERTOLLI, E.; NETO, J. P. D. Sentinel node status is the most important prognostic information for clinical stage IIB and IIC melanoma patients. **Annals of Surgical Oncology**, v. 27, n. 11, p. 4133–4140, 2020. Citation on page 24.

GANDER, W.; GAUTSCHI, W. Adaptive quadrature—revisited. **BIT Numerical Mathematics**, v. 40, n. 1, p. 84–101, 2000. Citation on page 32.

GERSHENWALD, J. E.; SCOLYER, R. A.; HESS, K. R.; SONDAK, V. K.; LONG, G. V.; ROSS, M. I.; LAZAR, A. J.; FARIES, M. B.; KIRKWOOD, J. M.; MCARTHUR, G. A. *et al.* Melanoma staging: Evidence-based changes in the american joint committee on cancer eighth edition cancer staging manual. **CA: A Cancer Journal for Clinicians**, v. 67, p. 472–492, 2017. Citations on pages 22 and 51.

GRAMBSCH, P. M.; THERNEAU, T. M. Proportional hazards tests and diagnostics based on weighted residuals. **Biometrika**, Biometrika Trust, v. 81, p. 515–526, 1994. Citations on pages 19, 28, and 44.

HESS, K. R. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. **Statistics in Medicine**, v. 14, p. 1707–1723, 1995. Citation on page 20.

HOUGAARD, P. Survival models for heterogeneous populations derived from stable distributions. **Biometrika**, v. 73, p. 387–396, 1986. Citation on page 35.

____. Modelling heterogeneity in survival data. **Journal of Applied Probability**, v. 28, p. 695–701, 1991. Citation on page 20.

____. Frailty models for survival data. **Lifetime Data Analysis**, Springer, v. 1, p. 255–273, 1995. Citation on page 20.

HOUGAARD, P.; MYGLEGAARD, P.; BORCH-JOHNSEN, K. Heterogeneity models of disease susceptibility, with application to diabetic nephropathy. **Biometrics**, v. 50, p. 1178–1188, 1994. Citation on page 20.

IARC. **Cancer Today**. available at gco.iarc.fr. Citation on page 19.

KALBFLEISCH, J. D.; PRENTICE, R. L. **The statistical analysis of failure time data**. [S.l.]: John Wiley & Sons, 2011. Citation on page 20.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, v. 53, p. 457–481, 1958. Citations on pages 22 and 27.

KLEIN, J. P.; MOESCHBERGER, M. L. Survival analysis: Statistical methods for censored and truncated data. **Springer Verlag, New York**, 2003. Citations on pages 19, 24, 25, 27, and 28.

LOUZADA-NETO, F. Extended hazard regression model for reliability and survival analysis. **Lifetime Data Analysis**, v. 3, p. 367–381, 1997. Citation on page 20.

____. Polyhazard models for lifetime data. **Biometrics**, v. 55, p. 1281–1285, 1999. Citation on page 20.

MACKENZIE, G. Regression models for survival data: the generalized time-dependent logistic family. **The Statistician**, JSTOR, v. 45, p. 21–34, 1996. Citation on page 20.

MALLER, R.; ZHOU, X. **Survival Analysis with Long-Term Survivors**. [S.l.]: John Wiley & Sons, 1996. Citation on page 49.

MILANI, E. A.; DINIZ, C. A. R.; TOMAZELLA, V. L. Generalized time-dependent complement log-log model. **Chilean Journal of Statistics**, v. 5, n. 1, p. 29–44, 2014. Citations on pages 24, 29, 30, 35, 37, 39, 47, and 53.

MILANI, E. A.; TOMAZELLA, V. L.; DIAS, T. C.; LOUZADA, F. *et al.* 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. Citation on page 20.

MOLINA, K. C.; CALSAVARA, V. F.; TOMAZELLA, V.; MILANI, E. A. Survival models induced by zero-modified power series discrete frailty: Application with a melanoma data set. **Statistical Methods in Medical Research**, p. 1–16, 2021. Citations on pages 44 and 50.

OAKES, D. A model for association in bivariate survival data. **Journal of the Royal Statistical Society B**, v. 44, p. 414–422, 1982. Citation on page 20.

PENG, Y.; TAYLOR, J. M. G.; YU, B. A marginal regression model for multivariate failure time data with a surviving fraction. **Lifetime Data Analysis**, v. 13, p. 351–369, 2007. Citation on page 20.

PETTITT, A.; DAUD, I. B. Investigating time dependence in Cox's proportional hazards model. **Applied Statistics**, JSTOR, v. 39, p. 313–329, 1990. Citations on pages 19 and 28.

PRENTICE, R. L. Linear rank tests with right censored data. **Biometrika**, v. 65, p. 167–179, 1978. Citation on page 20.

PRICE, D. L.; MANATUNGA, A. K. Modelling survival data with a cured fraction using frailty models. **Statistics in Medicine**, v. 20, p. 1515–1527, 2001. Citation on page 20.

PUZA, C. J.; BRESSLER, E. S.; TERANDO, A. M.; HOWARD, J. H.; BROWN, M. C.; HANKS, B.; SALAMA, A. K.; BEASLEY, G. M. The emerging role of surgery for patients with advanced melanoma treated with immunotherapy. **Journal of Surgical Research**, v. 236, p. 209–215, 2019. Citation on page 44.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Citations on pages 31, 39, and 40.

ROCHA, R.; NADARAJAH, S.; TOMAZELLA, V.; LOUZADA, F. Two new defective distributions based on the Marshall–Olkin extension. **Lifetime Data Analysis**, v. 22, p. 216–240, 2016. Citation on page 21.

____. A new class of defective models based on the Marshall–Olkin family of distributions for cure rate modeling. **Computational Statistics & Data Analysis**, Elsevier, v. 107, p. 48–63, 2017. Citation on page 21.

ROCHA, R.; NADARAJAH, S.; TOMAZELLA, V.; LOUZADA, F.; EUDES, A. New defective models based on the Kumaraswamy family of distributions with application to cancer data sets. **Statistical Methods in Medical Research**, v. 26, p. 1737–1755, 2017. Citation on page 21.

RODRIGUES, A. S.; CALSAVARA, V. F.; BERTOLLI, E.; PERES, S. V.; TOMAZELLA, V. L. Bayesian long-term survival model including a frailty term: Application to melanoma data. **Chilean Journal of Statistics**, v. 12, n. 1, p. 53–69, 2021. Citations on pages 44 and 50.

SABEL, M. S.; GRIFFITH, K.; SONDAK, V. K.; LOWE, L.; SCHWARTZ, J. L.; CIMMINO, V. M.; CHANG, A. E.; REES, R. S.; BRADFORD, C. R.; JOHNSON, T. M. Predictors of nonsentinel lymph node positivity in patients with a positive sentinel node for melanoma. **Journal of the American College of Surgeons**, v. 201, p. 37–47, 2005. Citation on page 51.

SCHEMPER, M. Cox analysis of survival data with non-proportional hazard functions. **The Statistician**, JSTOR, v. 41, p. 455–465, 1992. Citations on pages 19 and 28.

SCHOENFELD, D. Partial residuals for the proportional hazards regression model. **Biometrika**, v. 69, p. 239–241, 1982. Citations on pages 19 and 28.

SCUDILIO, J.; CALSAVARA, V. F.; ROCHA, R.; LOUZADA, F.; TOMAZELLA, V.; RODRIGUES, A. S. Defective models induced by gamma frailty term for survival data with cured fraction. **Journal of Applied Statistics**, Taylor & Francis, v. 46, p. 484–507, 2019. Citation on page 21.

SINHA, D.; DEY, D. Semiparametric Bayesian analysis of survival data. **Journal of the American Statistical Association**, American Statistical Association, v. 92, p. 1195–1212, 1997. Citation on page 20.

TWEEDIE, M. C. K. An index which distinguishes between some important exponential families. In: **Statistics: Applications and New Directions: Proc. Indian Statistical Institute Golden Jubilee International Conference**. [S.l.: s.n.], 1984. p. 579–604. Citation on page 35.

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. Citations on pages 20 and 35.

WIENKE, A. **Frailty Models in Survival Analysis**. Boca Raton: Chapman & Hall/CRC, 2011. Citations on pages 35 and 36.

YU, B.; PENG, Y. Mixture cure models for multivariate survival data. **Computational Statistics & Data Analysis**, v. 52, p. 1524–1532, 2008. Citation on page 20.

ICMC USP
SÃO CARLOS