UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modeling survival data based on a reparameterized weighted Lindley distribution

Alex Leal Mota

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

Alex Leal Mota

Modeling survival data based on a reparameterized weighted Lindley distribution

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Francisco Louzada Neto
Co-advisor: Prof. Dr. Vera Lucia D. Tomazella

USP – São Carlos
August 2022

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

M917m

Mota, Alex Leal
Modeling survival data based on a reparameterized weighted Lindley distribution / Alex Leal Mota; orientador Francisco Louzada Neto; coorientadora Vera Lucia Damasceno Tomazella. -- São Carlos, 2022.
129 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Cure fraction. 2. Frailty model. 3. GTDL model. 4. Maximum likelihood method. 5. Reparameterized weighted Lindley distribution. I. Neto, Francisco Louzada , orient. II. Tomazella, Vera Lucia Damasceno, coorient. III. 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

Alex Leal Mota

Modelagem de dados de sobrevivência baseada em uma distribuição de Lindley ponderada reparametrizada

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutor em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Francisco Louzada Neto
Coorientador: Prof. Dr. Vera Lucia D. Tomazella

USP – São Carlos
Agosto de 2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Alex Leal Mota, realizada em 24/06/2022.

Comissão Julgadora:

Prof. Dr. Francisco Louzada Neto (USP)

Prof. Dr. Manoel Ferreira dos Santos Neto (UFCG)

Prof. Dr. Eder Angelo Milani (UFG)

Profa. Dra. Lia Hanna Martins Morita (UFMT)

Prof. Dr. Jeremias da Silva Leão (UFAM)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

To my family

ACKNOWLEDGEMENTS

To God for life, health, wisdom, protection and for all the blessings given to me.

To my supervisors, Prof. Dr. Francisco Louzada Neto and Prof. Dr. Vera Lucia Damasceno Tomazella, for the trust, incentives, and important contributions to the elaboration and conclusion of this work.

To Prof. Dr. Eder A. Milani, Prof. Dr. Manoel Santos-Neto, Prof. Dr. Jeremias Leão, and Prof. Dr. Lia Hanna Morita for their important comments and suggestions for improving the text.

To all my family, especially my mother Lucilene Leal Mota, and my brothers Marcos André, Daniel Renato, Maria Eliene and Antônio Neto, who, with great affection and support, spared no effort for me to reach this stage of my life.

To all my friends and colleagues of PIPGES for the partnership, help, and moments of relaxation during the course. Special thanks go to Milena Lima, Marcílio Cardial, Isaac Olmos, Milton Miranda, Osafu Egbon, Asrat Belachew, Oilson Gonzatto, Marcos Jardel, Josimara Tatiane, Thiago Melo, Katy Molina, and Jonathan Vasquez.

To all my coauthors for the collaboration in the development of the articles that make up this work.

To all the members of the Petrobras project for sharing their experiences in various research topics.

To Mr. Manoel Silveira and his wife Mrs. Suely Silveira for the excellent accommodation in São Carlos-SP during the course.

Finally, my thanks go to CAPES and Petrobras for the financial support during the years of studies.

“All models are wrong, but some are useful.”
(George E. P. Box, 1976)

RESUMO

MOTA, ALEX LEAL. **Modelagem de dados de sobrevivência baseada em uma distribuição de Lindley ponderada reparametrizada**. 2022. 129 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Neste trabalho, propomos diferentes modelagens estatísticas para dados de sobrevivência baseadas em uma distribuição de Lindley ponderada reparametrizada. Inicialmente, apresentamos esta distribuição e estudamos suas propriedades matemáticas, estimação de máxima verossimilhança e simulações numéricas. Em seguida, propomos um novo modelo de fragilidade usando a distribuição de Lindley ponderada reparametrizada para modelar a heterogeneidade não observada em dados de sobrevivência univariados. A fragilidade é introduzida multiplicativamente na função de risco de base. Obtemos as funções de sobrevivência e risco não condicionais através da função transformada de Laplace da distribuição de fragilidade. Assumimos as funções de risco das distribuições Weibull e Gompertz como as funções de risco de base e usamos o método de máxima verossimilhança para estimar os parâmetros dos modelos resultantes. Estudos de simulação são realizados para verificar o comportamento dos estimadores propostos sob diferentes proporções de censura à direita e para avaliar o desempenho do teste da razão de verossimilhança para detectar heterogeneidade não observada em diferentes tamanhos amostrais. Além disso, propomos um modelo de longa duração com fragilidade Lindley ponderada reparametrizada. Uma vantagem do modelo proposto é modelar conjuntamente a heterogeneidade entre os pacientes por suas fragilidades e a presença de uma fração curada. Assumimos que o número desconhecido de causas competitivas que podem influenciar o tempo de sobrevivência segue uma distribuição binomial negativa e que o tempo para a k-ésima causa competitiva produzir o evento de interesse segue o modelo de fragilidade de Lindley ponderado reparametrizado com distribuição de base de Weibull. Alguns casos especiais do modelo são apresentados e a fração de cura é modelada usando a função de ligação logit. Novamente, usamos o método de máxima verossimilhança sob censura aleatória à direita para estimar os parâmetros do modelo proposto. Além disso, apresentamos estudos de simulação de Monte Carlo para verificar o comportamento dos estimadores de máxima verossimilhança assumindo diferentes tamanhos de amostra e proporções de censura. Finalmente, estendemos o modelo de regressão logística generalizado dependente do tempo incorporando fragilidades de Lindley ponderadas reparametrizadas. Essa modelagem proposta possui várias características importantes, tais como riscos não proporcionais, identifica a presença de sobreviventes de longa duração sem a adição de novos parâmetros, captura a heterogeneidade não observada, permite a interseção de curvas de sobrevivência e permite função de risco decrescente ou unimodal. Novamente, a estimação de parâmetros é realizada usando o método de máxima verossimilhança. Estudos de simulação de Monte Carlo são conduzidos para avaliar as propriedades assintóticas dos estimadores, bem

como algumas propriedades do modelo. A potencialidade de todos os modelos propostos é analisada empregando conjuntos de dados reais e comparações de modelos são realizadas.

Palavras-chave: distribuição Lindley ponderada reparametrizada, fração de cura, máxima verossimilhança, modelo de fragilidade, modelo logístico generalizado dependente do tempo, riscos não proporcionais.

ABSTRACT

MOTA, ALEX LEAL. **Modeling survival data based on a reparameterized weighted Lindley distribution**. 2022. 129 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

In this work, we propose different statistical modeling for survival data based on a reparameterized weighted Lindley distribution. Initially, we present this distribution and study its mathematical properties, maximum likelihood estimation, and numerical simulations. Then, we propose a novel frailty model by using the reparameterized weighted Lindley distribution for modeling unobserved heterogeneity in univariate survival data. The frailty is introduced multiplicatively on the baseline hazard function. We obtain unconditional survival and hazard functions through the Laplace transform function of the frailty distribution. We assume hazard functions of the Weibull and Gompertz distributions as the baseline hazard functions and use the maximum likelihood method for estimating the resulting model parameters. Simulation studies are further performed to verify the behavior of maximum likelihood estimators under different proportions of right-censoring and to assess the performance of the likelihood ratio test to detect unobserved heterogeneity in different sample sizes. Also, we propose a frailty long-term model where the frailties are described by reparameterized weighted Lindley distribution. An advantage of the proposed model is to jointly model the heterogeneity among patients by their frailties and the presence of a cured fraction of them. We assume that the unknown number of competing causes that can influence the survival time follows a negative binomial distribution and that the time for the k-th competing cause to produce the event of interest follows the reparameterized weighted Lindley frailty model with Weibull baseline distribution. Some special cases of the model are presented. The cure fraction is modeled by using the logit link function. Again, we use the maximum likelihood method under random right-censoring to estimate the proposed model parameters. Further, we present a Monte Carlo simulation study to verify the maximum likelihood estimators' behavior assuming different sample sizes and censoring proportions. Finally, we extend the non-proportional generalized time-dependent logistic regression model by incorporating reparameterized weighted Lindley frailties. This proposed modeling has several important characteristics, such as non-proportional hazards, identifies the presence of long-term survivors without the addition of new parameters, captures the unobserved heterogeneity, allows the intersection of survival curves, and allows decreasing or unimodal hazard function. Again, parameter estimation is performed using the maximum likelihood method. Monte Carlo simulation studies are conducted to evaluate the asymptotic properties of the estimators as well as some properties of the model. The potentiality of all the proposed models is analyzed by employing real datasets and model comparisons are performed.

Keywords: cure fraction, frailty model, generalized time-dependent logistic model, maximum likelihood method, non-proportional hazards, reparameterized weighted Lindley distribution.

LIST OF FIGURES

Figure 1 – Plots of PDF of the RWL distribution for different values of ϕ and μ fixed (top left and right panel and botton left panel), and RWL variance versus ϕ (botton right panel). 43

Figure 2 – Plots of PDF of the RWL distribution for different values of μ and ϕ fixed. . 44

Figure 3 – Plots of the hazard function of the RWL distribution. 45

Figure 4 – Plots of the MRL function of the RWL distribution. 48

Figure 5 – Left panel: Fitted survival functions superimposed to the estimated KM survival curve, considering the electrical appliances data. Right panel: Estimated hazard function of the RWL distribution for these data. 57

Figure 6 – Left panel: Fitted survival functions superimposed to the estimated KM survival function, considering the lifetimes of an agricultural machine. Right panel: Estimated hazard function of the RWL distribution for these data. . . 58

Figure 7 – Unconditional survival (left panel) and hazard (right panel) functions of the RWL frailty model with Weibull baseline hazard function. 64

Figure 8 – Unconditional survival (left panel) and hazard (right panel) functions of the RWL frailty model with Gompertz baseline hazard function for some parameter values. 65

Figure 9 – Estimated survival curve obtained via KM for the lung cancer dataset. . . . 74

Figure 10 – Exponential QQ plots of Cox-Snell residuals from the RWL frailty models with Weibull (left panel) and Gompertz (right panel) baseline hazard functions for the lung cancer dataset. 77

Figure 11 – Some shapes of PDF (right panel), survival (panel middle) and hazard (left panel) functions of the NBCrRWLF model. 82

Figure 12 – Some particular cases of the NBCrRWLF model. 83

Figure 13 – Empirical MRE, RMSE and 95% CP for the MLEs of θ, η, p_{00} and p_{01} from the PoCrRWLF model, under the indicated n, θ, η, p_{00} and p_{01} values, and also considering $p.cens = 0.35$. 87

Figure 14 – Empirical MRE, RMSE and 95% CP for the MLEs of θ, η, p_{00} and p_{01} from the PoCrRWLF model, under the indicated n, θ, η, p_{00} and p_{01} values, and also considering $p.cens = 0.75$. 88

Figure 15 – Empirical MRE, RMSE and 95% CP for the MLEs of θ, η, p_{00} and p_{01} from the PoCrRWLF model, under the indicated n, θ, η, p_{00} and p_{01} values, and also considering $p.cens = 0.85$. 88

Figure 16 – Estimated survival curve obtained via KM for the stomach cancer data set. . 89
Figure 17 – Estimated survival curve obtained via KM, adopting the clinical stage variable. 90
Figure 18 – First row, left to right: Estimated survival curve obtained via KM (solid lines) and BerCrRWLF and PoCrRWLF models (dashed lines), respectively. Second row, left to right: Estimated survival curve obtained via KM (solid lines) and GeoCrRWLF and NBCrRWLF models (dashed lines), respectively. 92
Figure 19 – Plot of normal theoretical quantiles *versus* quantile residuals considering the PoCrRWLF model. 93
Figure 20 – Unconditional survival (left panel) and hazard (right panel) functions from the GTDL-RWLF model. 97
Figure 21 – Estimated parameter biases associated to fitted incomplete model considering different parameter α_2 and β_2 values. 106
Figure 22 – Estimated survival curve obtained via KM for the lung cancer dataset. . . . 108
Figure 23 – Plot of log cumulative baseline hazard versus time of follow-up for the gender, age, surgery, clinical stage, radiotherapy and chemoterapy. 108
Figure 24 – Estimated survival curve obtained via KM (full line) for lung cancer dataset, and estimated survival function according to GTDL-RWLF model (dashed line) for gender, age at diagnosis, surgery, clinical stage, radiotherapy and chemotherapy. 112

LIST OF TABLES

Table 1 – MRE, MSE, CP and expected censoring proportion estimates for $N = 10,000$ samples of sizes $n \in \{20,50,100,200,400\}$, with 0%, 25% and 50% of random censored data, for $\mu = 0.5$ and $\phi = 0.7$. 53

Table 2 – MRE, MSE, CP and expected censoring proportion estimates for $N = 10,000$ samples of sizes $n \in \{20,50,100,200,400\}$, with 0%, 25% and 50% of random censored data, for $\mu = 2$ and $\phi = 5$. 53

Table 3 – Number of cycles, divided by 1,000, up to the failure for 60 electrical appliances in a life test. 55

Table 4 – MLEs, SEs and 95% CIs for the parameters of the RWL distribution, considering the electrical appliances data. 55

Table 5 – Model selection criteria values and KS test (statistic and p-values) for the fitted probability distributions, considering the electrical appliances data. . . 56

Table 6 – Times up to corrective maintenance of an agricultural machine ("+" denotes censoring). 56

Table 7 – MLEs, SEs and 95% CIs for the parameters of the RWL distribution, considering the agricultural machine data. 57

Table 8 – Model selection criteria values for the fitted probability distributions, considering the agricultural machine data. 58

Table 9 – Days to perform preventive maintenance to agricultural machine by assuming different percentages of failures, based on the RWL distribution. 59

Table 10 – Bias, RMSEs and SDs of the MLEs, and empirical CPs of 95% asymptotic CIs for the simulated data of the RWL frailty model with Weibull baseline hazard function 70

Table 11 – Bias, RMSEs and SDs of the MLEs, and empirical CPs of 95% asymptotic CIs for the simulated data of the RWL frailty model with Gompertz baseline hazard function. 71

Table 12 – Rejection rates of the null hypothesis (absence of unobservable heterogeneity) at 5% nominal significance level for several unobserved heterogeneity and sample sizes considering the RWL frailty model with Weibull baseline hazard function. 72

Table 13 – Rejection rates of the null hypothesis (absence of unobservable heterogeneity) at 5% nominal significance level for several unobserved heterogeneity and sample sizes considering the RWL frailty model with Gompertz baseline hazard function. 72
Table 14 – Descriptive analysis of the observed covariates from the lung cancer dataset. . 74
Table 15 – MLEs, SEs, information criteria and twice the logarithm of BF for the fitted frailty and Cox PH models considering the lung cancer dataset. The BF values were calculated assuming the RWL frailty models as correct. 76
Table 16 – Estimated specific lung survival rates for older patients stratified by gender, clinical stage and treatment under RWL frailty model. 78
Table 17 – MLE, SE, 95% asymptotic CIs, and AIC value obtained for the PoCrRWLF, BerCrRWLF, GeoCrRWLF and NBCrRWLF models considering clinical stage fitted to the stomach cancer data. 91
Table 18 – Bias, RMSE, SD, MSE, and CP of ML estimates for simulated data considering the GTDL-RWLF model for the scenario (i). 101
Table 19 – Bias, RMSE, SD, MSE, and CP of ML estimates for simulated data considering the GTDL-RWLF model for the scenario (ii). 102
Table 20 – Bias, RMSE, SD, MSE, and CP of ML estimates for simulated data considering the GTDL-RWLF model for the scenario (iii). 103
Table 21 – Percentage of the number of cases correctly identified by the proposed model when there was long-term survivors in a subgroup. 104
Table 22 – Percentage of the number of cases identified with PH using the proposed GTDL-RWLF. 104
Table 23 – Descriptive analysis of the observed covariates from the lung cancer dataset. . 107
Table 24 – Test of PHs assumption. 109
Table 25 – MLEs, SEs, 95% CIs, AIC value obtained for the traditional GTDL and GTDL-RWLF models considering gender, age at diagnosis, surgery, clinical stage, radiotherapy and chemotherapy fitted for the lung cancer dataset. . . . 111
Table 26 – MLEs, SEs, 95% asymptotic CIs, AIC value obtained for the traditional GTDL and GTDL-RWLF models considering gender, age at diagnosis, surgery, clinical stage, radiotherapy and chemotherapy fitted for the lung dataset. . . . 113
Table 27 – Estimated specific survival rates at 0.5-, 1-, 2- and 10-year for the GTDL-RWLF model considering clinical stage, surgery, gender, radiotherapy and chemotherapy for older patients. 114

CONTENTS

1 INTRODUCTION 21
1.1 Introduction and bibliographical review 21
1.2 Objectives of the thesis 26
1.3 Organization of the chapters 27
1.4 Products of the thesis 27

2 BACKGROUND 29
2.1 Basic concepts in survival analysis 29
2.2 Original WL distribution 32
2.3 Cox PH model 34
2.4 GTDL regression model: a NPH model 35
2.5 An unified version of the long-term survival models 37
2.6 Frailty models 38
2.6.1 Unconditional survival and hazard functions *39*

3 A REPARAMETERIZED WEIGHTED LINDLEY DISTRIBUTION: PROPERTIES, ESTIMATION AND APPLICATIONS 41
3.1 RWL distribution 41
3.2 Further properties of the RWL distribution 44
3.2.1 Quantile function *44*
3.2.2 Moments *45*
3.2.3 Mean residual life function *47*
3.2.4 Mean and median deviations *48*
3.2.5 Laplace transform *50*
3.3 Estimation 50
3.4 Results based on computation 52
3.5 Real data examples 54
3.5.1 Cycles up to the failure for electrical appliances *55*
3.5.2 Agricultural machine data *56*
3.6 Concluding remarks 59

4 A WEIGHTED LINDLEY FRAILTY MODEL: ESTIMATION AND APPLICATION TO A LUNG CANCER DATASET 61
4.1 RWL frailty model 62

4.1.1 ***RWL frailty model with Weibull baseline hazard function*** ***63***
4.1.2 ***RWL frailty model with Gompertz baseline hazard function*** ***63***
4.2 **Inference methods** 65
4.3 **Simulation study** 67
4.3.1 ***Asymptotic properties*** ***68***
4.3.2 ***Hypothesis testing*** $H_0 : \theta = 0$ ***69***
4.4 **Application on lung cancer data** 72
4.5 **Concluding remarks** 77

5 **A LONG-TERM FRAILTY REGRESSION MODEL BASED ON RWL DISTRIBUTION APPLIED TO STOMACH CANCER DATA** 79
5.1 **Model formulation** 79
5.1.1 ***Special cases of the NBCrRWLF model*** ***81***
5.1.1.1 *Bernoulli cure rate RWL frailty model* *81*
5.1.1.2 *Poisson cure rate RWL frailty model* *82*
5.1.1.3 *Geometric cure rate RWL frailty model* *83*
5.1.2 ***Inference methods*** ***84***
5.2 **Simulation study** 85
5.3 **Application on stomach cancer patients** 87
5.4 **Concluding remarks** 92

6 **NON-PROPORTIONAL HAZARDS MODEL WITH A FRAILTY TERM FOR MODELING SUBGROUPS WITH OR WITHOUT EVIDENCE OF LONG-TERM SURVIVORS** 95
6.1 **Model formulation** 95
6.2 **Inference** 97
6.3 **Simulation study** 99
6.3.1 ***Asymptotic properties*** ***100***
6.3.2 ***Sensibility analysis to detect long-term survivors in a subgroup*** . . . ***101***
6.3.3 ***Sensibility analysis to detect PH*** ***103***
6.3.4 ***Analysis of bias caused by the absence of a covariate*** ***105***
6.4 **Application on lung cancer data** 105
6.5 **Concluding remarks** 112

7 **FINAL REMARKS** 115

BIBLIOGRAPHY 117

CHAPTER

1

INTRODUCTION

1.1 Introduction and bibliographical review

Survival analysis plays an important role in medicine, epidemiology, biology, demography, economics, engineering, actuarial science, and other fields. It has expanded rapidly in the last three decades, with works having been published in various disciplines in addition to statistics. But what distinguishes survival analysis from other fields of statistics? Why does survival data need a special statistical theory? The main problem is censoring, which means that, for some individuals in the study population, the researcher only has the information that the event of interest did not occur before a particular time point. To put it plainly, a censored observation contains only partial information about the random variable of interest. Therefore, this kind of incomplete observation needs special methods (WIENKE, 2010).

In statistical literature, several parametric and non-parametric methods are available for modeling survival data (COLOSIMO; GIOLO, 2006; WIENKE, 2010; LAWLESS, 2011). The Lindley distribution is a one-parameter lifetime distribution which was originally proposed by Lindley (1958) in the context of Fiducial and Bayesian Statistics. Ghitany, Atieh and Nadarajah (2008) studied its mathematical properties, such as moments, hazard, mean residual life, entropy function, and asymptotic distribution of the extreme order statistics and inferential procedures. Moreover, the authors showed that such distribution outperforms the exponential model in many situations, which allowed its application in several real problems. However, the Lindley distribution supports solely increasing hazard function, and hence it does not provide enough flexibility for analyzing different types of lifetime data. Therefore, in recent years, many generalizations based on Lindley distribution have been proposed in the literature in order to increase its flexibility for modeling purposes. For example, generalized Lindley (ZAKERZADEH; DOLATI, 2009), weighted Lindley (GHITANY *et al.*, 2011), extended Lindley (BAKOUCH *et al.*, 2012), Power Lindley (GHITANY *et al.*, 2013), Transmuted two-parameter Lindley (KEMALOGLU; YILMAZ, 2017), Weibull Lindley (ASGHARZADEH; NADARAJAH; SHARAFI, 2018), Weibull

Marshall–Olkin Lindley (AFIFY *et al.*, 2020), among other distributions.

The weighted Lindley (WL) distribution has become much popular due to its simplicity, attractive properties, and flexibility to fit data when compared with similar generalizations of the exponential model, such as gamma and Weibull, among others. The use of the WL distribution is particularly appealing because: (i) its probability density function (PDF) takes decreasing, unimodal, or decreasing-increasing-decreasing shapes according to its shape parameter values. Note that most classical two-parameter distributions such as gamma, Weibull, Lognormal, and Gompertz distributions have either decreasing or unimodal densities. In contrast, the PDF of the WL model adds an extra shape that can be useful for modeling bimodal data; (ii) its hazard function presents monotone (increasing) and non-monotone (bathtub) shapes, thereby making this distribution compatible with modeling biological data from mortality studies. A few of the two-parameter lifetime distributions are capable of modeling the data exhibiting the bathtub-shaped hazard function. Here, we can cite the Wilson-Hilferty (WILSON; HILFERTY, 1931; RAMOS *et al.*, 2019), Nakagami-m (NAKAGAMI, 1960), exponential power (SMITH; BAIN, 1975), and Chen (CHEN, 2000) distributions; (iii) finally, the WL distribution can be written as a two-component mixture of gamma distributions, which facilitates obtaining attractive statistical properties including moments, survival, hazard, mean residual life, and characteristic functions, stochastic ordering, Bonferroni and the Lorenz curves, various entropies, order statistics derivations among other properties (GHITANY *et al.*, 2011; ALI, 2015) as well as efficient generation of random samples (MAZUCHELI *et al.*, 2016).

Al-Mutairi, Ghitany and Kundu (2015) estimated the stress-strength parameter $R = P(Y < X)$ when X and Y are two independent WL random variables with a common shape parameter. Mazucheli, Coelho-Barros and Louzada (2016) studied the Type I error rate and power for various tests such as likelihood ratio, Wald, modified Wald, Score and Gradient used to distinguish the WL distribution from basic Lindley distribution. By means of simulation studies, the authors have concluded that likelihood ratio and Score tests perform better than others with respect to size and power, respectively. Louzada and Ramos (2017) developed a long-term WL model by using the standard mixture cure rate model (BERKSON; GAGE, 1952) and studied its properties. Ghitany and Wang (2019) showed that as one of the shape parameters becomes large (smaller) the WL distribution can be approximated by a normal distribution (exponential, respectively). Bourguignon (2019) used the WL distribution as a conjugate prior probability distribution for Poisson and normal (with mean known) distributions. Besides, the author argued that the WL distribution can be used as a conjugate prior probability distribution to many other likelihood distributions such as the exponential, Pareto (with a known minimum), gamma (with the known shape), inverse gamma (with known shape parameter), lognormal (with known mean), Weibull (with the known shape) and inverse Gaussian (with known mean). Accordingly, the WL distribution of conjugate priors stands for a more flexible class of priors than the class of gamma prior distributions. Classical and Bayesian estimation methods for the shape parameters of the WL distribution can be found in Mazucheli, Louzada and Ghitany (2013), Al-Zahrani

and Gindwan (2014), Ali (2015), Ramos, Louzada and Cancho (2017), Ghitany, Song and Wang (2017), and Kim and Jang (2021). Some generalizations of the WL distribution are due to Asgharzadeh *et al.* (2016), Ramos and Louzada (2016), and Shanker, Shukla and Leonida (2019).

Models based on the hazard function became remarkable in survival analysis since the construction of the Cox proportional hazards (PH) model (COX, 1972). According to Wienke (2010), one of the reasons this model is so popular is the ease of dealing with technical troubles such as censoring and truncation. This is due to the appealing interpretation of the hazard as a risk that changes over time. Naturally, the concept allows for the entering of covariates in order to describe their influence and to model different risk levels for different subgroups. In this model, the PH assumption states that the hazard ratio for two subjects who are characterized by different sets of covariates depends only on the values of these covariates and does not depend on time. In other words: the hazard ratio is constant over time which means that the effect of a given covariate on a hazard level is the same at all time (BORUCKA, 2014). However, in practice, the PH assumption is restrictive and for various reasons, non-proportional hazards (NPH) are often observed in many studies. For example, in medical studies, the most common types of NPH are time-dependent treatment effects, delayed treatment effects, crossing hazards, and diminishing treatment effects over time (FISHER; LIN, 1999; LIN *et al.*, 2020; PHINYO; PATUMANOND; PONGUDOM, 2021). In such scenarios, the use of the Cox PH model, in its original form, is not adequate (HOSMER; LEMESHOW, 1999; BOX-STEFFENSMEIER; ZORN, 2001; COLOSIMO; GIOLO, 2006). Nevertheless, Schemper (1992) has noted that the Cox model has undoubtedly been used in many cases in which proportionality assumptions are violated, negatively impacting the results.

Modifications and extensions of Cox PH model to deal with NPH have been proposed by several authors (SCHEMPER, 1992; HASTIE; TIBSHIRANI, 1993; KLEINBAUM; KLEIN, 2012a; KLEINBAUM; KLEIN, 2012b; BORUCKA, 2014; RATNANINGSIH; SAEFUDDIN; KURNIA, 2021). Although this, there is no “natural", widely accepted approach, and obtaining a satisfactory model can be complicated. Furthermore, there are further concerns about the complexity involved in the practical interpretation of the coefficients and in the robustness of such models. Therefore, alternative NPH models have also been introduced in the statistical literature. Aranda-Ordaz (1983) as well as Tibshirani and Ciampi (1983) have proposed proportional and additive hazards models for grouped data. Thomas (1986) and Sasieni (1996) introduced excess hazards models. Aalen (1980) developed additive hazard model and McKeague and Sasieni (1994) presented a partly parametric version of it. Prentice (1978) proposed the accelerated failure time (AFT) model, which considers that the covariates have a multiplicative effect both on time and on the baseline hazard function and, hence, NPH situations are accommodated. Etezadi-Amoli and Ciampi (1987) presented the extended hazard regression model, which is currently known as *hybrid hazard model* since Cox’s PH and AFT models come up as special cases. Generalizations of the hybrid hazards model have been made to allow the spread parameter

to be dependent on covariates and to deal with latent competing risks (LOUZADA-NETO, 1997; LOUZADA-NETO, 1999).

Another NPH model is the so-called generalized time-dependent logistic (GTDL) regression model. The GTDL regression model was introduced by Mackenzie (1996) as an alternative to the Cox PH model. In this model, the time effect is captured by one of its parameters, allowing NPH functions. However, when the time effect goes to zero, the GTDL model tends to be a PH model. Accordingly, the GTDL is flexible enough to model PH and non-PH survival data. An advantage of the GTDL model over the NPH models aforementioned is that its survival function can be improper, that is, $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(t) = p_0$, where $p_0 \in (0,1)$. For censored data, this is a desirable property since in many studies the longest failure times tend to be censored, and hence the empirical survivor function does not fall to 0 (MACKENZIE, 1996). For example, in medical studies, $p_0 \in (0,1)$ can be interpreted as a cure fraction or long-term survivors proportion in the study population (MALLER; ZHOU, 1996; IBRAHIM; CHEN; SINHA, 2001; KLEIN; MOESCHBERGER, 2003). A Bayesian approach for estimating the GTDL model parameters based on Markov Chain Monte Carlo (MCMC) methods can be found in (LOUZADA-NETO; CREMASCO; MACKENZIE, 2010). Extensions of the GTDL model have been made by Milani *et al.* (2015) and Calsavara *et al.* (2019a) in order to give more flexibility in the fitting of the univariate survival data. While earlier Ha and MacKenzie (2010) introduced a multivariate version from the GTDL regression model to model multivariate (or correlated) survival data.

Survival models with cure fraction often referred to as cure rate models (or long-term models) have been used for modeling time-to-event data for various types of cancers, including breast, non-Hodgkins lymphoma, leukemia, prostate, melanoma, and head and neck cancers, where for these diseases, a significant proportion of patients are “cured" due to, for example, by a genetic predisposition or a treatment (IBRAHIM; CHEN; SINHA, 2001). The two most widely applied cure rate models are the standard mixture model (SMM) firstly proposed by Boag (1949) and modified by Berkson and Gage (1952), and the promotion time model by Yakovlev, Tsodikov and Bass (1993). These two approaches differ in how they deal with the distribution of the latent number of causes of the event of interest. In the cure rate modeling by using the SMM, the unknown number of competing causes is supposed to be a Bernoulli random variable distributed, whereas, in the promotion time modeling, this number follows a Poisson distribution. As pointed out by Ortega *et al.* (2015), in a biological context, the idea behind these assumptions lies within a latent competing cause structure, in the sense that the event of interest can be a tumor recurrence or the death of a patient, occurring due to unknown competing causes. According to Ibrahim, Chen and Sinha (2001), these latent competing causes possibly are assigned to metastasis-competent tumor cells left active after initial treatment, such as radiotherapy, chemotherapy, surgery, among others. A metastasis-competent tumor cell is a tumor cell having the potential of metastasizing (TSODIKOV; YAKOVLEV; ASSELAIN, 1996). If tumor recurrence or death did not occur, one can consider the patient to be cured. For a more

detailed review of this and other cure rate models, the interested reader is referred to (MALLER; ZHOU, 1996; IBRAHIM; CHEN; SINHA, 2001; TSODIKOV; IBRAHIM; YAKOVLEV, 2003).

Rodrigues *et al.* (2009) proposed an unification of long-term survival models. In this approach, the number of competing causes of the event of interest is assumed to follow any positive discrete distribution possessing a probability generating function (PGF) (FELLER, 2008). This approach extends and unifies the long-term survival models proposed by Berkson and Gage (1952), Yakovlev, Tsodikov and Bass (1993) and Chen, Ibrahim and Sinha (1999). Many cure rate models proposed in last years have been formulated by using different distributions for the number of competing causes. Examples include the negative binomial (RODRIGUES *et al.*, 2009), COM-Poisson (RODRIGUES *et al.*, 2011), power series (ORTEGA *et al.*, 2015), Yule-Simon (GALLARDO; GÓMEZ; BOLFARINE, 2017), polylogarithm (GALLARDO; GÓMEZ; CASTRO, 2018), modified power series (GALLARDO *et al.*, 2020) , zero-inflated power series (CANCHO *et al.*, 2020), zero-modified geometric (LEÃO *et al.*, 2020), Waring (VASQUEZ; RODRIGUES; BALAKRISHNAN, 2020) and Bell (GALLARDO; CASTRO; GÓMEZ, 2021) distributions, among many others.

In practice, random effects or unexplained heterogeneity are very often present in survival data, perhaps almost always (AALEN, 1988). Hence, it is useful to consider two sources of heterogeneity in survival data: observed heterogeneity accounted for by observable risk factors included in the model (and therefore theoretically predictable) and unobserved heterogeneity caused by random effects being theoretically unpredictable. According to Hougaard (1991), there are advantages in considering these two sources of heterogeneity separately: unobserved heterogeneity may explain some unexpected results or gives an alternative explanation for some aspects as, for example, nonproportional or decreasing hazard functions. Some authors such as Struthers and Kalbfleisch (1986), Bretagnolle and Huber-Carol (1988), and Henderson and Oman (1999) have investigated the effects of ignoring unobserved heterogeneity. These studies concluded that biased regression estimates were obtained, which should not be used. As an alternative, frailty models must be considered in analysis (DUCHATEAU; JANSSEN, 2007; WIENKE, 2010; HOUGAARD, 2012). In these models, an unobservable random effect (termed by Vaupel, Manton and Stallard (1979) as *"frailty"*) is introduced on the baseline hazard function to control for unobservable heterogeneity among subjects under study. Consequently, it is expected that the subjects who are most frail will have a higher risk, and consequently, they will experience the event of interest sooner than those who are less frail (WIENKE, 2010).

Due to the randomness of the frailty, it is commonly modeled by a probability distribution usually referred to as *frailty distribution*. Even though a non-parametric specification of the frailty distribution is possible (HOROWITZ, 1999; ALMEIDA *et al.*, 2020), the parametric approach is often employed by a question of mathematical convenience, as pointed out by Wienke (2010). In addition, its variability determines the degree of unobserved heterogeneity and indicates the inadequacy of the baseline model considered. Vaupel, Manton and Stallard

(1979) first reported the use of gamma distribution as a standard assumption for frailty, which continues to be used currently. This assumption is mainly employed because it provides an easy mathematical treatment for obtaining analytical expressions for unconditional survival and hazard functions by using the Laplace transform. Hence, traditional maximum likelihood (ML) methods can be used for estimating the model parameters (WIENKE, 2010). However, other frailty distributions have also been proposed as an alternative to the gamma distribution (HOUGAARD, 1995). For instance, uniform, Weibull, lognormal, positive stable, inverse Gaussian (IG), and compound Poisson distributions, in addition to the power variance function (PVF), which includes most of these models (VAUPEL; YASHIN, 1983; HOUGAARD, 1986; HOUGAARD, 1995; AALEN, 1988). Wienke (2010) and Hougaard (2012) compared most of these frailty distributions and discussed their specific advantages and limitations in applications to real data examples. Generalized gamma and Birnbaum-Saunders (BS) distributions have recently been introduced as frailty distributions by Balakrishnan and Peng (2006) and Leão *et al.* (2017), respectively. The generalized gamma distribution as frailty distribution has similar characteristics as the PVF, which can also provide an excellent fit to the data. However, it has no closed-form expression for the unconditional survival function, and more sophisticated estimation strategies as numerical integration and Monte Carlo simulation are required, which can make the use of the PVF preferable in practice. Finally, though the BS frailty distribution has a Laplace transform mathematically tractable, its variance is limited. Hence, it should not be used in applications with more significant unobserved heterogeneity (LOUZADA *et al.*, 2020).

1.2 Objectives of the thesis

In recent years, some traditional distributions have been reparameterized in terms of its mean and/or precision parameters to model real problems; see, e.g., Ferrari and Cribari-Neto (2004), Cepeda and Gamerman (2005), Santos-Neto *et al.* (2016), Rigby *et al.* (2019), Bourguignon and Gallardo (2020), Gallardo *et al.* (2020), and Bourguignon, Santos-Neto and Castro (2021). Some advantages of using reparameterized distributions in statistical modeling are: (i) they simplify the classical and Bayesian inferences; (ii) they facility the interpretation of results; (iii) they allow us to model heteroscedasticity (in regression); (iv) in survival analysis, they can be an alternative to existing frailty distributions (LEÃO *et al.*, 2018; LEÃO *et al.*, 2017). Following this approach, Mazucheli, Coelho-Barros and Achcar (2016) introduced an alternative parameterization for the WL distribution in the context of orthogonal parameters (COX; REID, 1987), which we will call reparameterized WL (RWL) distribution throughout the thesis. Orthogonal parameters have many advantages in the inference results as, for example, for large sample sizes we have independence among the maximum likelihood of the orthogonal parameters, since the Fisher information matrix is diagonal. Other advantages of orthogonal parameters can be found in (COX; REID, 1987).

The general objective of this thesis is to propose different statistical modeling for survival

data based on the RWL distribution. Some specific objectives are:

- to study several mathematical properties of the RWL distribution and propose maximum likelihood estimation assuming the presence of censored and uncensored data;
- to propose a RWL frailty model as an alternative to the existing frailty models for modeling unobserved heterogeneity in univariate survival data;
- to develop a new long-term frailty regression model based on the RWL distribution to jointly account for the heterogeneity among patients by their frailties and the presence of a cured fraction of them.
- to introduce a new extended version of the GTDL regression model by incorporating RWL frailty in order to model survival data under non-proportional hazards, different time effects in groups, cure fractions in one or both groups, and unobserved heterogeneity in study population.

1.3 Organization of the chapters

The remainder of this thesis is organized as follows. In Chapter 2, we briefly presented a background about survival analysis, genesis of the WL distribution and its main properties, Cox PH, GTDL regression, cure rate, and frailty models. In Chapter 3, we discussed several properties of the RWL distribution and propose maximum likelihood estimation assuming the presence of censored and uncensored data. In Chapter 4, we introduce the RWL frailty model for modeling unobserved heterogeneity in univariate survival data, while in Chapter 5 we developed a new long-term frailty regression model based on the RWL distribution to jointly model the heterogeneity among patients by their frailties and the presence of a cured fraction of them. In Chapter 6, we consider the GTDL regression model with a RWL frailty term. Such a methodology extends the GTDL model and identifies several important characteristics, such as non-proportional hazards; identifying the presence of long-term survivors without the addition of new parameters; capturing the unobserved heterogeneity (if present in the dataset); allowing the intersection of survival curves, as well as decreasing and unimodal hazard functions. Finally, we present a discussion, conclusions, and future research in Chapter 7.

1.4 Products of the thesis

- Mota, A., Ramos, P. L., Ferreira, P., Tomazella, V., & Louzada, F. (2021). A Reparameterized Weighted Lindley Distribution: Properties, Estimation and Applications. *Revista Colombiana de Estadística*, 44(1), 65-90. <http://dx.doi.org/10.15446/rce.v44n1.86566>

- Mota, A., Milani, E. A., Calsavara, V. F., Tomazella, V. L., Leão, J., Ramos, P. L., Ferreira, Paulo H., & Louzada, F. (2021). Weighted Lindley frailty model: estimation and application to lung cancer data. *Lifetime Data Analysis*, 1-27. <https://doi.org/10.1007/s10985-021-09529-1>

- Mota, A.L., Milani, Eder A., Leão, Jeremias, Ramos, Pedro L., Ferreira, Paulo H., Gonzatto, Oilson, Tomazella, Vera L. D., & Louzada, Francisco. (2021). A new long-term frailty regression model based on a weighted Lindley distribution applied to stomach cancer data (under review).

- Gazon, A. B., Milani, E. A., Mota, A. L., Louzada, F., Tomazella, V. D., & Calsavara, V. F. (2021). Nonproportional hazards model with a frailty term for modeling subgroups with evidence of long-term survivors: Application to a lung cancer dataset. *Biometrical Journal*, 63(6), 1–26. <https://doi.org/10.1002/bimj.202000292>

CHAPTER

2

BACKGROUND

In this chapter, we present some basic concepts in survival analysis such as censoring schemes, main functions used in this area and its mathematical relationships. We also present briefly the original parameterization of the WL distribution and its main properties. The Cox PH and GTDL models as well as the unified version of cure rate or long-term models proposed by Rodrigues *et al.* (2009) are also discussed. Finally, we describe about frailty models and discuss how to obtain the unconditional survival and hazard functions.

2.1 Basic concepts in survival analysis

Survival analysis is the collection of statistical procedures for data analysis in which the response variable is time until an event of interest occurs (DAVID; MITCHEL, 2012). This event of interest may be death, the appearance of a tumor, remission after some treatment, equipment breakdown, divorce, cessation of smoking, and so forth. Depending on the type of application, survival analysis is also known as lifetime data analysis, reliability analysis, time to event analysis, and event history analysis. Hence, the response variable is also referred to as survival time, failure time, lifetime, risk period, event time, time to event, and duration time (WIENKE, 2010).

A key characteristic that distinguishes survival analysis from other areas in statistics is that survival data are usually censored. Censoring is present when we have some information about a subject's event time, but we don't know the exact event time. According to Klein and Moeschberger (2006), possible censoring schemes are:

- right censoring: all that is known is that the individual is still alive at a given time;
- left censoring: all that is known is that the individual has experienced the event of interest prior to the start of the study;

- interval censoring: the only information is that the event occurs within some interval.

The right-censored observations can be Type I censoring, Type II censoring, or random censoring. Type I censoring occurs when the event is observed only if it occurs prior to some prespecified time. Type II censoring (often used in engineering) occurs when the study continues until the failure of the first r subjects, where r is some predetermined integer $(1 \leq r < n)$. Finally, random censoring occurs when the subject leaves the study without having experienced the event of interest, or when the subject may experience some competing event which causes them to be removed from the study. In practice, random censoring is the most commonly censoring type. Moreover, Type I and Type II right-censoring mechanisms can be seen as particular cases of random censoring (COLOSIMO; GIOLO, 2006). Regardless of the type of censoring, we must assume that it is non-informative about the event; that is, the censoring is caused by something other than the impending failure.

Let T be a continuous nonnegative random variable representing the time from a well-defined specific starting point until the occurrence of an event. The distribution of the random variable T can be specified through mathematically related functions, where the knowledge of one of them is sufficient to derive the others. These functions are called PDF, survival and hazard functions and are particularly useful in survival applications.

The PDF, denoted by $f(t)$, is defined as limit of the probability that an individual fails in the short interval $[t, t+\Delta t]$, per unit width Δt, or simply as the probability of failure in a small interval per unit time (LEE; WANG, 2003). Mathematically, the PDF is a continuous function at t given by

$$f(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T \leq t+\Delta t]}{\Delta t}, \quad t > 0, \tag{2.1}$$

where $f(t) \geq 0$ and $\int_0^\infty f(t) = 1$.

The survival function is defined as the probability of an individual does not fail until the time t, that is, the probability of an individual's surviving till time t. Mathematically, the survival function is expressed as

$$S(t) = P[T \geq t] = \int_t^\infty f(u)du, \quad t > 0, \tag{2.2}$$

where $f(\cdot)$ is the corresponding PDF. Note that the survival function can be also expressed as $S(t) = 1 - F(t)$, where $F(t)$ is the cumulative distribution function, defined by $F(t) = P[T \leq t] = \int_0^t f(u)du$. Hence, the cumulative distribution function represents the probability of an individual fails before t. The survival function, $S(t)$, is a nonicreasing continuous function of time t with properties: (i) $S(0) = 1$ and (ii) $S(\infty) = \lim_{t\to\infty} S(t) = 0$.

According to Wienke (2010), the major concept in survival analysis is the hazard function. This function is also called (depending on the field of application), hazard rate, mortality rate,

incidence rate, mortality curve, failure rate, or force of mortality. The hazard function, denoted by $h(t)$, is the instantaneous failure rate in time t, given survival up to time t and is defined by,

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T \leq t + \Delta t | T \geq t]}{\Delta t}, \quad t > 0. \tag{2.3}$$

The hazard function has been preferable by many authors to describe the behavior of survival time due to its interpretation. Moreover, according to its behavior, the hazard function can character special classes of the survival time distributions. for that reason, the modelling of the hazard function is an important method for survival data analysis.

Another useful function in survival data analysis is called cumulative hazard function, denoted by $H(t)$, defined by

$$H(t) = \int_0^t h(u)du, \quad t > 0, \tag{2.4}$$

where $h(\cdot)$ is the hazard function. Even though the cumulative hazard function does not has an direct interpretation, it is quite useful in the evaluation of the hazard function, specially in nonparametric estimation, where it is possible to find an estimator with great properties for cumulative hazard function, whereas the hazard function is difficult to be estimated (NELSON, 1972; AALEN, 1978; COLOSIMO; GIOLO, 2006).

Below we derive some useful relationships from these functions:

1. By definitions of the PDF, cumulative distribution function (CDF) and survival function, we get

$$f(t) = \frac{d}{dt}F(t) \Rightarrow f(t) = -\frac{d}{dt}S(t). \tag{2.5}$$

2. By definition of the hazard function, we have

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P[t \leq T \leq t + \Delta t | T \geq t]}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P[t \leq T \leq t + \Delta t]}{\Delta t P[T \geq t]} \\ &= \lim_{\Delta t \to 0} \left[\frac{F(t + \Delta t) - F(t)}{\Delta t}\right] \frac{1}{S(t)}. \end{aligned}$$

By derivative definition (STEWART, 2015),

$$h(t) = \left[\frac{d}{dt}F(t)\right] \frac{1}{S(t)},$$

and from (2.5), we obtain

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.6}$$

3. From (2.5) and (2.6), we get

$$h(t) = -\frac{\frac{d}{dt}S(t)}{S(t)} = -\frac{d}{dt}\log S(t). \tag{2.7}$$

4. Now integrating both sides of (2.7), and then exponentiating, we are led

$$S(t) = \exp\left(-\int_0^t h(u)du\right) = \exp\left(-H(t)\right). \tag{2.8}$$

Since $S(\infty) = 0$, it follows from (2.8) that $H(\infty) = \lim_{t\to\infty} H(t) = \infty$. Hence, the hazard function, $h(t)$, has the properties

$$h(t) \geq 0 \text{ and } \int_0^\infty h(t)dt = \infty.$$

5. Finally, in addition to (2.8), if follows from (2.6) that

$$f(t) = h(t)S(t) = h(t)\exp\left(-H(t)\right).$$

2.2 Original WL distribution

When an investigator records an observation by nature according to certain stochastic model, the recorded observation will not have the original distribution unless every observation is given an equal chance of being recorded. For example, suppose that the original observation t_0 comes from a distribution with PDF $f_0(t_0 \mid \theta_1)$, where θ_1 is a parameter vector, and that observation t is recorded according to a probability re-weighted by a weight function $\omega(t \mid \theta_2) > 0$, with θ_2 being a new parameter vector, then t comes from a distribution with PDF

$$f(t \mid \theta) = A\omega(t \mid \theta_2)f_0(t \mid \theta_1), \tag{2.9}$$

where $\theta = (\theta_1, \theta_2)$ and A is a normalizing constant. Distributions of this type are called weighted distributions. This class of distributions was firstly introduced by Rao (1965) and a survey of their applications can be found in (PATIL *et al.*, 1977). When $\omega(t \mid \theta_2)$ is a constant, we have $f(t \mid \theta) = f_0(t \mid \theta_1)$. On the other hand, if $\omega(t \mid \theta_2) = t$, the resulting weighted distribution is called length-bias distribution (PATIL; RAO, 1978).

The class of the weighted distributions provides a new understanding of standard distributions as well as methods of extending distributions in order to add more flexibility in fitting data. Taking into account this class, Ghitany *et al.* (2011) developed a two-parameter WL distribution as a generalization of the lifetime Lindley distribution (LINDLEY, 1958). Let $\omega(t \mid \theta_2) = t^{\phi-1}$, for $\phi > 0$, and consider that $f_0(t \mid \lambda)$ is the PDF of the one-parameter Lindley distribution (LINDLEY, 1958), whose PDF is given by

$$f_0(t \mid \lambda) = \frac{\lambda^2}{\lambda+1}(1+t)\exp\left(-\lambda t\right), \quad \lambda, t > 0. \tag{2.10}$$

Then, from (2.9) and (2.10), the class of WL distributions have PDFs given by

$$f(t \mid \lambda, \phi) = Bt^{\phi-1}(1+t)\exp(-\lambda t), \quad t > 0, \tag{2.11}$$

where B is a constant normalizing.

The PDF of the WL distribution introduced by Ghitany *et al.* (2011) is expressed as

$$f(t \mid \lambda, \phi) = \frac{\lambda^{\phi+1}}{(\lambda+\phi)\Gamma(\phi)} t^{\phi-1}(1+t)\exp(-\lambda t), \quad t > 0, \tag{2.12}$$

where $\lambda > 0$ and $\phi > 0$ are shape parameters, and $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ is the gamma function. Note that when $\phi = 1$, the WL distribution reduces to Lindley distribution (LINDLEY, 1958). On the other hand, as $\phi \to \infty$ ($\phi \to 0$) the WL distribution can be approximated by normal distribution with mean ϕ/λ and variance ϕ/λ^2 (standard exponential) (GHITANY; WANG, 2019).

According to Ghitany *et al.* (2011), the PDF (2.12) can be decreasing, unimodal, or decreasing-increasing-decreasing depending on the values selected for the parameters. Most classical two-parameter distributions such as Weibull, gamma and Gompertz distributions have either decreasing or unimodal densities. The PDF of the WL model adds an extra shape which can be useful for modeling bimodal data. In addition, the WL distribution can be written as a two-component mixture as follows

$$f(t \mid \lambda, \phi) = \left(\frac{\lambda}{\lambda+\phi}\right) f_1(t \mid \lambda, \phi) + \left(\frac{\phi}{\lambda+\phi}\right) f_2(t \mid \lambda, \phi), \tag{2.13}$$

where $f_j(t \mid \lambda, \phi)$ is the PDF of the gamma distribution, with shape parameter $\phi + j - 1$ and scale parameter λ, denoted by Gamma$(\phi + j - 1, \lambda)$, $j = 1, 2$. This representation facilitates obtaining properties of the WL distribution, since the properties of the gamma distribution are well-known in the statistical literature. For example, the mean and variance of the WL distribution are readily given by

$$\mathbb{E}[T] = \frac{\phi(\lambda+\phi+1)}{\lambda(\lambda+\phi)} \quad \text{and} \quad \mathbb{V}ar[T] = \frac{(\phi+1)(\lambda+\phi)^2 - \lambda^2}{\lambda^2(\lambda+\phi)^2},$$

respectively.

The cumulative distribution of the WL distribution is given by

$$F(t \mid \lambda, \phi) = \frac{(\lambda+\phi)\gamma(\phi, \lambda t) - (\lambda t)^\phi \exp(-\lambda t)}{(\lambda+\phi)\Gamma(\phi)}, \tag{2.14}$$

where $\gamma(a, b) = \int_0^b t^{a-1}\exp(-t)dt$, $a, b > 0$, is the lower incomplete gamma function.

From (2.14), the corresponding survival and hazard functions are expressed, respectively, as

$$S(t \mid \lambda, \phi) = \frac{(\lambda+\phi)\Gamma(\phi, \lambda t) + (\lambda t)^\phi \exp(-\lambda t)}{(\lambda+\phi)\Gamma(\phi)},$$

and

$$h(t \mid \lambda, \phi) = \frac{\lambda^{\phi+1} t^{\phi-1}(1+t)\exp(-\lambda t)}{(\lambda+\phi)\Gamma(\phi, \lambda t) + (\lambda t)^{\phi}\exp\{-\lambda t\}},$$

where $\Gamma(a,z) = \int_z^{\infty} t^{a-1}\exp(-t)dt$, $a > 0$ and $z \geq 0$, is the upper incomplete gamma function. The hazard function, $h(t)$, of the WL distribution is bathtub shaped (increasing) when $0 < \phi < 1$ ($\phi \geq 1$), for all $\lambda > 0$ (GHITANY *et al.*, 2011).

The mixture representation (2.13) is also useful to generate random samples from WL distribution. In fact, in this case, we can generate random variables using the Algorithm 1:

Algorithm 1 – Generator of random numbers from WL distribution.

1: Define the values of $\boldsymbol{\Theta} = (\lambda, \phi)^{\top}$;
2: Generate $u \sim \text{Uniform}(0,1)$;
3: If $u \leq \dfrac{\lambda}{\lambda+\phi}$, generate $t \sim \text{Gamma}(\phi, \lambda)$. Otherwise, generate $t \sim \text{Gamma}(\phi+1, \lambda)$;
4: Repeat the previous steps to obtain the desired sample size.

Fortunately, the `rwlindley()` function within the `LindleyR` package can be used for this purpose. In addition, this package computes the probability density, the cumulative distribution, the quantile, and the hazard functions and generates random deviates from the discrete and continuous Lindley distribution as well as for 19 of its modifications. It also generates censored random deviates from any probability distribution available in R; see (MAZUCHELI *et al.*, 2016; R Core Team, 2021) for more details.

2.3 Cox PH model

The Cox PH regression model (COX, 1972) is essentially one of the most commonly used models in survival analysis for investigating the association between the lifetimes of patients and one or more covariates. One of the main reasons for this popularity is the ease with which technical difficulties such as censoring and truncation are handled. This is due to the appealing interpretation of the hazard rate as a risk that changes over time (WIENKE, 2010). Another reason is because of the availability of easy-to-use software (HENDERSON; OMAN, 1999).

The idea of the Cox PH model is to define a hazard level as a dependent variable which is explained by the time-related component (so-called baseline hazard) and the covariates-related component. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^{\top}$ be a $p \times 1$ vector of covariates. The Cox PH model is given by

$$h(t|\boldsymbol{x}) = h_0(t)\exp\left(\boldsymbol{x}^{\top}\boldsymbol{\beta}\right), \tag{2.15}$$

where $\boldsymbol{\beta}$ is the $p \times 1$ vector, for $(p < n)$, of unknown regression coefficients associated with the covariates $\boldsymbol{x}$, and $h_0(\cdot)$ is called baseline hazard function because $h(t|\boldsymbol{x}) = h_0(t)$ when $\boldsymbol{x} = 0$. The

exponential form $\exp\left(\boldsymbol{x}^\top\boldsymbol{\beta}\right)$ in (2.15) is usually used for convenience. If desired, $\exp\left(\boldsymbol{x}^\top\boldsymbol{\beta}\right)$ can be replaced with some other nonnegative function $g(\boldsymbol{x}^\top\boldsymbol{\beta})$ such that $g(\mathbf{0}) = 1$.

The baseline hazard function, $h_0(t)$, can be choice of two ways: non-parametrically or parametrically. If the baseline hazard function is choice be non-parametric, then the Cox's PH model is termed semiparametric PH model, since $\exp(\boldsymbol{x}^\top\boldsymbol{\beta})$ is a parametric component. In this case, the model parameters are estimated by using the partial likelihood function (COX, 1975). On the other hand, when the baseline hazard function is parametric, the Cox's PH model is termed fully parametric PH model or simply parametric PH model. Hence, traditional likelihood methods for censored or uncensored data can be used for estimating the model parameters; see (COLOSIMO; GIOLO, 2006; KLEIN; MOESCHBERGER, 2006; KALBFLEISCH; PRENTICE, 2011) for details.

The key assumption in the Cox regression model is of PH, which means that the hazard ratio is constant over time, or that the hazard for an individual is proportional to the hazard for any other individual. Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be the covariates for two individuals, then the ratio hazards is given by

$$\frac{h(t|\boldsymbol{x}_1)}{h(t|\boldsymbol{x}_2)} = \frac{h_0(t)\exp\left(\boldsymbol{x}_1^\top\boldsymbol{\beta}\right)}{h_0(t)\exp\left(\boldsymbol{x}_2^\top\boldsymbol{\beta}\right)} = \frac{\exp\left(\boldsymbol{x}_1^\top\boldsymbol{\beta}\right)}{\exp\left(\boldsymbol{x}_2^\top\boldsymbol{\beta}\right)} = \exp\left\{(\boldsymbol{x}_1-\boldsymbol{x}_2)^\top\boldsymbol{\beta}\right\}, \tag{2.16}$$

which is a constant, independent of time. Hence, if an individual at the beginning of the study has a hazard equal to twice the hazard of another individual, this hazard ratio will be the same for the entire follow-up period (COLOSIMO; GIOLO, 2006). Therefore, the Cox's model must not be used in situations where the non-proportionally of hazards is evident. However, Schemper (1992) have noted that the Cox model has undoubtedly been used in many cases in which proportionality assumptions are violated, with consequences for the results. In order to detect the violations of proportionally of hazards, we can use tests of proportionally, Schoenfeld residuals and graphical methods; see (KLEIN; MOESCHBERGER, 2006; SCHOENFELD, 1982; HESS, 1995).

2.4 GTDL regression model: a NPH model

In practice, one usually fits a Cox PH model and assesses the proportionality assumption. According to Calsavara *et al.* (2019a), when departures from assumption are detected, several possible workarounds, such as redefinition of covariates, model stratification by a covariate with a non-proportional hazard, use of time dependent covariate terms, use of separate models for disjunct time periods, and fitting of a NPH model, can be applied. This latter approach is becoming increasingly in analyzing survival data; see (AALEN, 1980; KALBFLEISCH; PRENTICE, 2011; ETEZADI-AMOLI; CIAMPI, 1987) and references cited therein. In this context, Mackenzie (1996) proposed a parametric NPHs model called GTDL regression model, whose hazard function is written as

$$h_0(t \mid \lambda, \alpha, \boldsymbol{\beta}) = \frac{\lambda\exp(\alpha t + \boldsymbol{x}_1^\top\boldsymbol{\beta})}{1+\exp(\alpha t + \boldsymbol{x}_1^\top\boldsymbol{\beta})}, \tag{2.17}$$

where $\lambda > 0$ is a scalar, $\alpha \in \mathbb{R}$ is a measure of the time effect, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a vector of p unknown parameters measuring the effect of the p covariates $\boldsymbol{x}_1 = (\boldsymbol{x}_{1_1}, \ldots, \boldsymbol{x}_{1_p})^\top$.

The ratio for the hazard functions of two individuals with different covariates is given by

$$\rho(\boldsymbol{x}_{1_1}, \boldsymbol{x}_{1_2}) = \frac{h_0(t \mid \lambda, \alpha, \boldsymbol{\beta}, \boldsymbol{x}_{1_1})}{h_0(t \mid \lambda, \alpha, \boldsymbol{\beta}, \boldsymbol{x}_{1_2})} = \frac{1 + \exp(\alpha t + \boldsymbol{x}_{1_2}^\top \boldsymbol{\beta})}{1 + \exp(\alpha t + \boldsymbol{x}_{1_1}^\top \boldsymbol{\beta})} \exp[(\boldsymbol{x}_{1_1} - \boldsymbol{x}_{1_2})^\top \boldsymbol{\beta}].$$

Note that the time effect does not disappear, so the non-proportionality becomes evident. Mackenzie (1996) argued that when $\left[1 + \exp(\alpha t + \boldsymbol{x}_1^\top \boldsymbol{\beta})\right]^{-1} \approx 1$, the GTDL model becomes a PH model and gives similar estimates to the Cox PH model. However, note that when the time effect tends to zero, the hazard ratios tend to be constant over time. Accordingly, the GTDL is flexible enough to model PH and non-PH survival data.

The respective cumulative hazard function is

$$H_0(t \mid \lambda, \alpha, \boldsymbol{\beta}) = \frac{\lambda}{\alpha} \log \left[\frac{1 + \exp(\alpha t + \boldsymbol{x}_1^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_1^\top \boldsymbol{\beta})} \right], \tag{2.18}$$

and the survival function is expressed by

$$S_0(t \mid \lambda, \alpha, \boldsymbol{\beta}) = \left[\frac{1 + \exp(\alpha t + \boldsymbol{x}_1^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_1^\top \boldsymbol{\beta})} \right]^{-\lambda/\alpha}. \tag{2.19}$$

The behavior of the hazard function (2.17) takes several forms according to the value of α: for $\alpha > 0$, the hazard function is increasing; for $\alpha < 0$, it is decreasing; and for $\alpha = 0$, this function is constant. The survival function (6.1) also has its behavior determined by the value of α. For $\alpha \geq 0$, it is proper, i.e, $S_0(0 \mid \lambda, \alpha, \boldsymbol{\beta}) = 1$ and $\lim_{t \to \infty} S_0(t \mid \lambda, \alpha, \boldsymbol{\beta}) = 0$; while for $\alpha < 0$, it is improper because $S_0(0 \mid \lambda, \alpha, \boldsymbol{\beta}) = 1$, but

$$\lim_{t \to \infty} S_0(t \mid \lambda, \alpha, \boldsymbol{\beta}) = p(\boldsymbol{x}_1) = \{1 + \exp(\boldsymbol{x}_1^\top \boldsymbol{\beta})\}^{\lambda/\alpha} \in (0, 1),$$

and, hence $p(\boldsymbol{x}_1)$ can be interpreted as the proportion of long-term survivors in the study population (MALLER; ZHOU, 1996).

An advantage of using the GTDL model is that it allows long-term survivors without needed of extra parameters. In addition, it does not make assumptions about the existence of long-term survivors in the study population. However, we notice that the time effect parameter, α, is unique for all groups, leaving the model biologically limited, since groups of patients who receive different treatments may have different time effects. Hence, a regression structure in this parameter turns the model more flexible. In addition, patients can be long-term survivors and the insertion of covariates in this parameter will reflect an estimate of $\alpha < 0$ (CALSAVARA *et al.*, 2019a).

2.5 An unified version of the long-term survival models

Rodrigues *et al.* (2009) proposed an unification of long-term survival models, which assume that the number of competing causes related of the event of interest follows any positive discrete distribution possessing a PGF (FELLER, 2008). We describe such a methodology as follows. For a subject in the population, let M be a positive unobserved discrete random variable denoting the number of competing causes related to the occurrence of an event of interest. For instance, in cancer studies M represents the number of carcinogenic cells at the end of treatment that can produce a detectable cancer. Assume that M has probability mass function defined by $p_m = P(M = m)$, $m = 0, 1, \ldots$. Given $M = m$, let W_k, $k = 1, \ldots, m$, be positive continuous random variables representing the time-to-event due to the k-th competing cause. We suppose that conditional on M, W_k's are independent and identically distributed (IID) random variables with distribution function $F(z) = 1 - S(z)$ that does not depend on M. The survival function $S(z)$ is proper, that is, $S(0) = 1$ and $S(\infty) = 0$. Hence, exponential, piecewise exponential, and Weibull distributions, for instance, can be used to represent W_k (IBRAHIM; CHEN; SINHA, 2001). Notwithstanding IID assumption on W_k's is surely strong, this option favors simplicity and analytical tractability at the expense of a more general formulation (CASTRO; CANCHO; RODRIGUES, 2009). In addition, notwithstanding this limitation, such models have proven to be useful in many real-world applications (ORTEGA *et al.*, 2015; LEÃO *et al.*, 2018; LEÃO *et al.*, 2020). The number of competing causes M and the lifetime W_k associated with a particular cause are not observable (latent variables). Therefore, the observed lifetime is defined as

$$T = \begin{cases} \min(W_1, W_2, \ldots, W_M), & \text{for} \quad M \geq 1; \\ +\infty, & \text{if} \quad M = 0; \end{cases} \tag{2.20}$$

which leads to a proportion of "cured" or "immune" individuals denoted by $p_0 = P(M = 0)$. Under this setup, the survival function for the population, $S_{pop}(t)$, is given by

$$\begin{aligned} S_{pop}(t) &= P(T \geq t) \\ &= P(M = 0) + P(W_1 \geq t, W_2 \geq t, \ldots, W_M \geq t, M \geq 1) \\ &= p_0 + \sum_{m=1}^{\infty} p_m P(W_1 \geq t, W_2 \geq t, \ldots, W_M \geq t \mid M = m) \\ &= \sum_{m=0}^{\infty} p_m [S(t)]^m \\ &= G_M(S(t)), \end{aligned} \tag{2.21}$$

where $G_M(\cdot)$ is the PGF of the random variable M, which converges when $S(t) \in [0, 1]$.

From (2.21), note that $S_{pop}(0) = 1$ and $S_{pop}(\infty) = \lim_{t \to \infty} S_{pop}(t) = p_0 \in (0, 1)$. Thus $S_{pop}(t)$ is an improper survival function and p_0 is the cured fraction that may be present in the population from which the data is taken (RODRIGUES *et al.*, 2009). The corresponding

improper PDF and hazard functions from (2.21) are given by

$$f_{pop}(t) = f(t)\left(\frac{d}{ds}G_M(s)\Big|_{s=S(t)}\right),$$

and

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} = \frac{f(t)\left(\frac{d}{ds}G_M(s)\Big|_{s=S(t)}\right)}{S_{pop}(t)},$$

respectively, where $f(t) = -\frac{d}{dt}S(t)$ is a proper PDF, that is, $f(t) \geq 0$ and $\int_{-\infty}^{\infty} f(t)dt = 1$.

Different discrete distributions have been employed for modeling the number of competing causes related to the occurrence of an event of interest. If the Bernoulli or the Poisson distribution are used, we obtain the SMM and promotion time cure rate models, respectively; see (BERKSON; GAGE, 1952; YAKOVLEV; TSODIKOV; BASS, 1993). Rodrigues *et al.* (2009) considered the negative binomial distribution which has the Bernoulli, binomial, Poisson and Geometric distributions as particular cases, Rodrigues *et al.* (2009) utilized the COMP-Poisson distribution, Ortega *et al.* (2015) employed the power series distribution, which has as special cases the Binomial, Poisson, Geometric, Negative Binomial, Logarithmic distributions, among others. Gallardo, Gómez and Bolfarine (2017) considered the Yale-Simon distribution, Leão *et al.* (2020) used the zero-modified geometric distribution, Vasquez, Rodrigues and Balakrishnan (2020) considered the Waring distribution, and Gallardo, Castro and Gómez (2021) employed the Bell distribution.

2.6 Frailty models

Frailty models can be used in survival analysis for modeling random effects or unexplained heterogeneity between individuals or groups (BARKER; HENDERSON, 2005). The idea of these models is to introduce a non-negative random effect (frailty) multiplicatively on the hazard function of the baseline model. As a consequence, subjects or groups which are most frail will have a higher risk, and hence, they will experience the event of interest (e.g., death, relapsed, or failure) sooner than those who are less frail (WIENKE, 2010).

Let $T > 0$ be a random variable representing the failure time and Z be an unobservable, non-negative random variable denoting the frailty. The conditional hazard function of T given $Z = z$ is expressed as

$$h(t \mid z) = zh_0(t), \tag{2.22}$$

where $h_0(\cdot)$ is the baseline hazard function which is assumed to be the same for all subjects Note that the frailty z factor acts multiplicatively on the baseline hazard function. Hence, frailty z

increases or decreases the risk of occurrence of the event of interest if $z > 1$ or $z < 1$, respectively. When $z = 1$ for all individuals, then the standard baseline model is obtained as a special case.

If some covariates are observed in the study, then they can be introduced in (2.22) of similar way to the Cox model (2.15). So that, the conditional hazard function of T given $Z = z$ become

$$h(t \mid z, \boldsymbol{x}) = z h_0(t) \exp(\boldsymbol{x}^\top \boldsymbol{\beta}), \tag{2.23}$$

where $\boldsymbol{\beta}$ is the $p \times 1$ vector, for $(p < n)$, of unknown regression coefficients associated with the $p \times 1$ vector of covariates $\boldsymbol{x}$.

As in Cox's model (2.15), the baseline hazard function, $h_0(\cdot)$, can be chosen non-parametrically or parametrically. In the non-parametric case, the Breslow and Nelson-Aalen estimators, as well as their modified versions are commonly used to estimate the cumulative hazard function. Hence, the model is known as a semiparametric frailty model, since it is assumed a parametric frailty distribution; see (NIELSEN *et al.*, 1992; KLEIN, 1992; PARNER *et al.*, 1998; BARKER; HENDERSON, 2005). On the other hand, in the parametric approach, baseline hazard functions of the distributions such as exponential, lognormal, Gompertz, and Weibull, among others, are often used (WIENKE, 2010).

The conditional survival function is obtained from (2.23) as follows

$$S(t \mid z, \boldsymbol{x}) = \exp\left[-z H_0(t) \exp(\boldsymbol{x}^\top \boldsymbol{\beta})\right], \tag{2.24}$$

where $H_0(t) = \int_0^t h_0(s) ds$ is the cumulative baseline hazard function.

2.6.1 Unconditional survival and hazard functions

In order to obtain unconditional survival and hazard functions, that is, not depending on unobserved quantities, we must integrate out the conditional survival function (2.24) on frailty. This is equivalent to calculate the Laplace transform of the frailty distribution. The definition of Laplace transform of a function is given as follows.

Definition 2.6.1. The Laplace transform of a function $f(y)$, for $y > 0$, at $s \in \mathbb{C}$, is the function $\mathscr{L}(s)$, which is defined by

$$\mathscr{L}_f(s) = \int_0^\infty \exp\{-sy\} f(y) dy.$$

Let $f(z)$ be the frailty PDF. According to Elbers and Ridder (1982), to satisfy the assumption of identifiability in resulting model, we need that the frailty distribution has mean one, that is, $\mathbb{E}[Z] = 1$. Thus, by integrating $S(t \mid z, \boldsymbol{x})$ given in (2.24) on $Z = z$, we obtain

$$S(t \mid \boldsymbol{x}) = \int_0^\infty \exp\left[-z H_0(t) \exp(\boldsymbol{x}^\top \boldsymbol{\beta})\right] f(z) dz = \mathscr{L}_f\left(H_0(t) \exp(\boldsymbol{x}^\top \boldsymbol{\beta})\right), \tag{2.25}$$

where $\mathscr{L}_f(\cdot)$ denotes the Laplace transform of the frailty distribution. Hence, the unconditional hazard function can be obtained from Equation (2.25) as:

$$h(t \mid \boldsymbol{x}) = -\frac{h_0(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta})\mathscr{L}_f'\left(H_0(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta})\right)}{\mathscr{L}_f\left(H_0(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta})\right)}, \tag{2.26}$$

where $\mathscr{L}_f'(t)$ is the first derivative of the Laplace transform with respect to time t. The unconditional survival and hazard functions (given above) measure, respectively, the survival and risk of an individual randomly drawn from a study population (WIENKE, 2010).

As noted above, the use of a frailty distribution that has a Laplace transform on the closed-form is essential for computing both unconditional survival and hazard functions, which simplifies parameter estimation. However, when the frailty distribution has no Laplace transform on the closed-form, numerical integration or Markov Chain Monte Carlo methods need to be applied (BALAKRISHNAN; PENG, 2006; HOUGAARD, 2012; ROBERT; CASELLA, 2013). In practice, computational convenience must be taken into account when considering frailty distribution in univariate and multivariate survival data modeling (PICKLES; CROUCHLEY, 1995; WIENKE, 2010).

CHAPTER

3

A REPARAMETERIZED WEIGHTED LINDLEY DISTRIBUTION: PROPERTIES, ESTIMATION AND APPLICATIONS

Although Mazucheli, Coelho-Barros and Achcar (2016) have proposed the RWL distribution, the authors did not study its properties and, in addition, they also did not consider the ML estimation for the parameters under censored data. Our objective in this chapter is to derive and discuss many mathematical properties of this distribution, including its moments, mean and median deviations, quantile, characteristic, hazard, mean residual life, and Laplace transform functions. Also, we show that the second parameter of this distribution can be interpreted as a precision parameter, which can be useful in further studies. The inference for the model parameters is conducted under the classical (or frequentist) framework via the ML method assuming the presence of uncensored and random censored data. Numerical simulations are carried out in order to investigate the performance of the ML estimators (MLEs) under different sample sizes and proportions of censored data. Finally, the applicability of the RWL distribution is illustrated in two real data sets.

3.1 RWL distribution

Mazucheli, Coelho-Barros and Achcar (2016) proposed a new parameterization of the WL distribution, which allows diverse features of data modeling to be considered. The RWL distribution is obtained by transforming (λ, ϕ) into (μ, ϕ), where

$$\mu = \frac{\phi(\lambda + \phi + 1)}{\lambda(\lambda + \phi)},$$

is the mean of the original parameterization (2.12). Hence, by inverse transformation, we obtain that

$$\lambda = \frac{\phi(1-\mu) + \sqrt{\phi^2(\mu-1)^2 + 4\mu\phi(\phi+1)}}{2\mu}.$$

Therefore, the PDF of the RWL distribution is expressed as

$$f(t \mid \mu, \phi) = \frac{[a(\mu,\phi)]^{\phi+1} t^{\phi-1}(1+t)\exp\left\{-\frac{a(\mu,\phi)}{2\mu}t\right\}}{(2\mu)^{\phi}\,[a(\mu,\phi)+2\mu\phi]\,\Gamma(\phi)}, \quad t>0, \tag{3.1}$$

where $a(\mu,\phi) = \phi(1-\mu) + \sqrt{\phi^2(\mu-1)^2 + 4\mu\phi(\phi+1)}$, $\mu > 0$ is the mean, that is, $\mathbb{E}[T] = \mu$ and $\phi > 0$ is the shape parameter. From now on, $T \sim \text{RWL}(\mu,\phi)$ will be used to denote that the random variable T follows this distribution.

As the original WL distribution, the PDF (3.1) can be written as a two mixture of the gamma distributions, that is,

$$f(t \mid \mu, \phi) = p f_1(t \mid \mu, \phi) + (1-p) f_2(t \mid \mu, \phi), \tag{3.2}$$

where $p = \dfrac{a(\mu,\phi)}{a(\mu,\phi)+2\mu\phi}$ and

$$f_j(t \mid \mu, \phi) = \left(\frac{a(\mu,\phi)}{2\mu}\right)^{\phi+j-1} \frac{t^{\phi+j-2}}{\Gamma(\phi+j-1)} \exp\left\{-\frac{a(\mu,\phi)}{2\mu}t\right\}, \quad t>0,$$

is the PDF of the gamma distribution with shape parameter $\phi + j - 1$ and scale parameter $a(\mu,\phi)/2\mu$, for $j = 1, 2$.

By using the mixture representation (3.2), we found that the variance of the PDF (3.1) is given by

$$\mathbb{V}\text{ar}[T] = \left(\frac{2\mu}{a(\mu,\phi)}\right)^2 \frac{\left[(\phi+1)\,(a(\mu,\phi)+2\mu\phi)^2 - a^2(\mu,\phi)\right]}{(a(\mu,\phi)+2\mu\phi)^2}. \tag{3.3}$$

Figure 1 shows shapes of the PDF (3.1) considering different values of ϕ, when μ is fixed, and RWL variance against ϕ. We note that the variability decreases when the shape parameter ϕ increases (see bottom right panel). Thus, ϕ can be interpreted as a precision parameter. On the other hand, when μ increases, the probabilities decrease and vice versa.

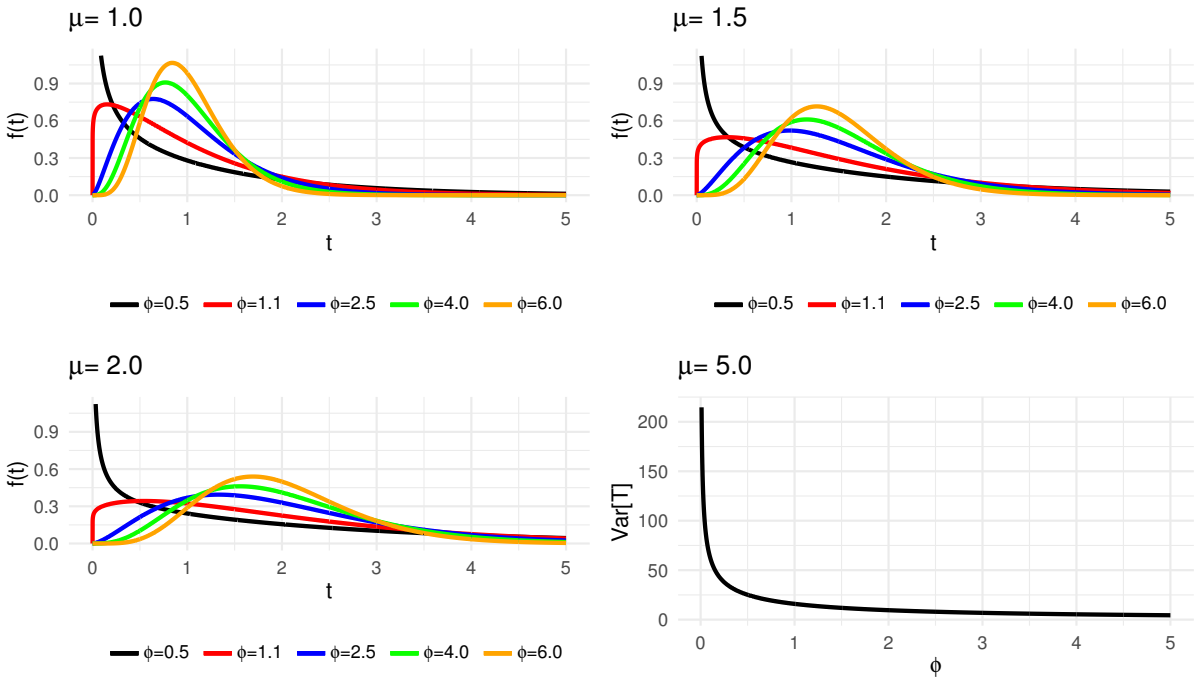

Figure 1 – Plots of PDF of the RWL distribution for different values of ϕ and μ fixed (top left and right panel and botton left panel), and RWL variance versus ϕ (botton right panel).

Source: Elaborated by the author.

Figure 2 displays shapes of the PDF (3.1) considering different values of μ, when ϕ is fixed. Note that when μ increases the distribution is more spread out and if μ decreases it becomes more concentrated around the mean. Thus, μ is the mean and also a scale parameter of this distribution. On the other hand, again we see that when ϕ is higher, the distribution has lower variability.

The corresponding survival and hazard functions of the RWL distribution are given, respectively, by

$$S(t \mid \mu,\phi) = \frac{1}{\Gamma(\phi)}\left[\Gamma\left(\phi, \frac{a(\mu,\phi)}{2\mu}t\right) + \frac{[a(\mu,\phi)t]^{\phi}\exp\left\{-\frac{a(\mu,\phi)}{2\mu}t\right\}}{(2\mu)^{\phi-1}[a(\mu,\phi)+2\mu\phi]}\right], \tag{3.4}$$

and

$$h(t \mid \mu,\phi) = \frac{[a(\mu,\phi)]^{\phi+1}t^{\phi-1}(1+t)\exp\left\{-\frac{a(\mu,\phi)}{2\mu}t\right\}}{2\mu\left[(2\mu)^{\phi-1}[a(\mu,\phi)+2\mu\phi]\Gamma\left(\phi, \frac{a(\mu,\phi)}{2\mu}t\right) + [a(\mu,\phi)t]^{\phi}\exp\left\{-\frac{a(\mu,\phi)}{2\mu}t\right\}\right]}, \tag{3.5}$$

where, for all $c > 0$ and $d \geq 0$,

$$\Gamma(c,d) = \int_d^{\infty} t^{c-1}e^{-t}dt,$$

is the upper incomplete gamma function. Since this function is widely available in computer programs, the practical use of Equations (3.4) and (3.5) present no problem; see the `zipfR` package within `R` software (EVERT; BARONI; EVERT, 2006).

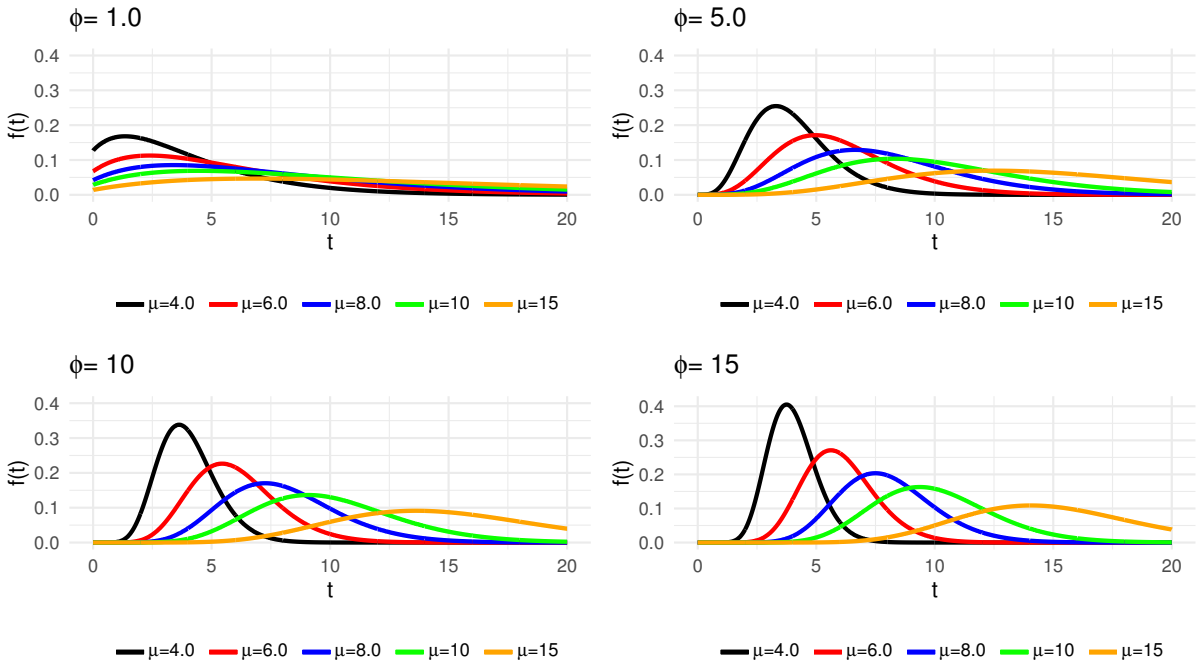

Figure 2 – Plots of PDF of the RWL distribution for different values of μ and ϕ fixed.

Source: Elaborated by the author.

Figure 3 shows different shapes for the hazard function of the RWL distribution, considering distinct values of μ and ϕ. It can be noted that the hazard function has monotonically increasing ($\phi \geq 1$) and bathtub ($\phi < 1$) shapes for all $\mu > 0$ (as the original WL distribution; see (GHITANY *et al.*, 2011)).

3.2 Further properties of the RWL distribution

In this section, we present some mathematical properties of the RWL distribution, such as r-th moments, characteristic function, and Laplace transform, among others.

3.2.1 Quantile function

The quantile function of a probability distribution is useful in statistical applications and Monte Carlo simulation. From (3.4), we have that the CDF of the RWL distribution is given by

$$F(t \mid \mu,\phi) = 1 - \frac{1}{\Gamma(\phi)}\left[\Gamma\left(\phi, \frac{a(\mu,\phi)}{2\mu}t\right) + \frac{[a(\mu,\phi)t]^{\phi}\exp\left\{-\frac{a(\mu,\phi)}{2\mu}t\right\}}{(2\mu)^{\phi-1}\left[a(\mu,\phi)+2\mu\phi\right]}\right].$$

Hence, the p-quantile, t_p, is obtained by solving the following equation:

$$\frac{[a(\mu,\phi)t_p]^{\phi}\exp\left\{-\frac{a(\mu,\phi)}{2\mu}t_p\right\}}{(2\mu)^{\phi-1}\left[a(\mu,\phi)+2\mu\phi\right]} = \Gamma(\phi)(1-p) - \Gamma\left(\phi, \frac{a(\mu,\phi)}{2\mu}t_p\right), \tag{3.6}$$

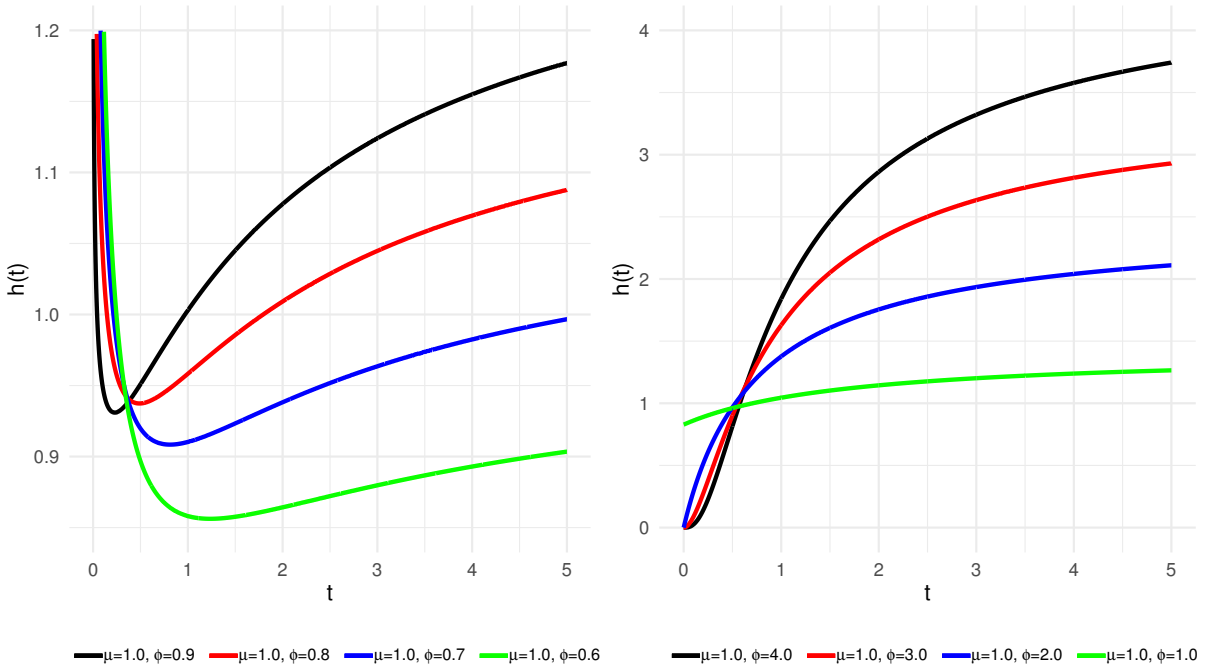

Figure 3 – Plots of the hazard function of the RWL distribution.

Source: Elaborated by the author.

for $0 < p < 1$. Observe that if $p = 0.5$ we get the median of the RWL distribution.

Note that the quantile function does not have a closed mathematical expression. In this case, the `uniroot` function of the `R` software can be used to find out the desired quantiles of the data; see (BRENT, 1973; R Core Team, 2021).

3.2.2 Moments

Many important characteristics and properties of a probability distribution can be obtained through its moments, such as mean, variance, skewness, and kurtosis.

Theorem 3.2.1. *If $T \sim RWL(\mu,\phi)$, then the r-th power, logarithmic and negative moments are given, respectively, by*

$$(i)\ \mathbb{E}[T^r] = \left[\frac{2\mu}{a(\mu,\phi)}\right]^r \frac{[a(\mu,\phi)+2\mu\phi+2\mu r]\Gamma(\phi+r)}{[a(\mu,\phi)+2\mu\phi]\Gamma(\phi)};$$

$$(ii)\ \mathbb{E}[\log(T^r)] = r\left[\psi(\phi)+\frac{2\mu}{a(\mu,\phi)+2\mu\phi}-\log\left(\frac{a(\mu,\phi)}{2\mu}\right)\right];$$

$$(iii)\ \mathbb{E}[T^{-r}] = \left[\frac{a(\mu,\phi)}{2\mu}\right]^r \frac{\Gamma(\phi-r)\,[a(\mu,\phi)+2\mu(\phi-r)]}{[a(\mu,\phi)+2\mu\phi]\Gamma(\phi)},$$

where $\psi(k) = \dfrac{d}{dk}\log(\Gamma(k))$ is the digamma function.

Proof. We will only prove the item (i) of Theorem 3.2.1 because the proof for the other remaining items follows similarly. In fact, let us use the mixture representation given in Equation (3.2). We then have

$$\mathbb{E}[T^r] = \left(\frac{a(\mu,\phi)}{a(\mu,\phi)+2\mu\phi}\right)\int_0^\infty t^r f_1(t \mid \mu,\phi)dt + \left(\frac{2\mu\phi}{a(\mu,\phi)+2\mu\phi}\right)\int_0^\infty t^r f_2(t \mid \mu,\phi)dt.$$

where $f_j(t \mid \mu,\phi)$ is the PDF of the gamma distribution with shape parameter $\phi+j-1$ and scale parameter $\frac{a(\mu,\phi)}{2\mu}$, for $j=1,2$. Note that

$$\int_0^\infty t^r f_j(t \mid \mu,\phi)dt = \left(\frac{2\mu}{a(\mu,\phi)}\right)^r \frac{\Gamma(\phi+j+r-1)}{\Gamma(\phi+j-1)}, \quad j=1,2.$$

Thus, after some algebraic manipulations, we finish the proof of this theorem. ■

The coefficient of variation (CV) is used to analyze the dispersion in terms of their average value when two or more data sets have different units of measure. As a result, we can say that the CV is a way of expressing the variability of the data, excluding the influence of the variable's order of magnitude. Often, the CV is given in percentage.

The CV of $T \sim \text{RWL}(\mu,\phi)$ is given by

$$CV[T] = \frac{\sqrt{Var[T]}}{\mathbb{E}[T]} = \sqrt{\frac{4\left[a(\mu,\phi)+2\mu\phi+4\mu\right]\phi(\phi+1)}{[a(\mu,\phi)]^2[a(\mu,\phi)+2\mu\phi]}-1}.$$

The next corollary gives us the harmonic mean of the RWL distribution. This measure of central tendency can be useful in many real problems; see, e.g., Hasna and Alouini (2004), Limbrunner, Vogel and Brown (2000) and Raftery *et al.* (2006).

Corollary 3.2.1. *The harmonic mean of the random variable $Y \sim RWL(\mu,\phi)$ is given by*

$$H_m = \left(\mathbb{E}\left[\frac{1}{T}\right]\right)^{-1} = \frac{2\mu\Gamma(\phi)[a(\mu,\phi)+2\mu\phi]}{a(\mu,\phi)\Gamma(\phi-1)[a(\mu,\phi)+2\mu(\phi-1)]}.$$

Proof. This result can be established by using the item (iii) of Theorem 3.2.1 with $r=1$ and then taking the reciprocal of the resulting expression. ■

Another way to characterize a distribution is by using its characteristic function (CF). The CF of a random variable is also known as Fourier transform of its PDF and has applications in the most diverse areas of scientific knowledge; see, e.g., Ingle, Kogon and Manolakis (2005), Yu (2004) and Lukacs (1972). Besides, the CF is also useful to compute the r-th power moments; see (BILLINGSLEY, 2008).

Theorem 3.2.2. *If $T \sim RWL(\mu,\phi)$, then its CF is given by*

$$\Psi_T(s) = \left(\frac{1}{a(\mu,\phi)+2\mu\phi}\right)\left(1-\frac{2\mu is}{a(\mu,\phi)}\right)^{-\phi}\left[a(\mu,\phi)+2\mu\phi\left(1-\frac{2\mu is}{a(\mu,\phi)}\right)^{-1}\right],$$

for all $s \in \mathbb{R}$, where $i=\sqrt{-1}$ is the imaginary unit.

Proof. In fact, by the representation of mixture given in Equation (3.2), we have

$$\begin{aligned}\Psi_T(s) &= \mathbb{E}[e^{isT}] = \int_0^\infty e^{ist} f(t \mid \mu,\phi)dt \\ &= \left(\frac{a(\mu,\phi)}{a(\mu,\phi)+\phi}\right)\int_0^\infty e^{ist} f_1(t \mid \mu,\phi)dt + \left(\frac{2\mu\phi}{a(\mu,\phi)+\phi}\right)\int_0^\infty e^{ist} f_2(t \mid \mu,\phi)dt.\end{aligned}$$

Now, as $f_j(t)$ is the PDF of gamma distribution with parameters $\phi+j-1$ and $a(\mu,\phi)/2\mu$, $j=1,2$, we then obtain

$$\Psi_T(s) = \left(\frac{a(\mu,\phi)}{a(\mu,\phi)+2\mu\phi}\right)\left(1-\frac{2\mu is}{a(\mu,\phi)}\right)^{-\phi} + \left(\frac{2\mu\phi}{a(\mu,\phi)+2\mu\phi}\right)\left(1-\frac{2\mu is}{a(\mu,\phi)}\right)^{-\phi-1}.$$

Now, after some algebraic manipulations, we get the desired result. ■

3.2.3 Mean residual life function

The mean residual life (MRL) function represents the expected additional lifetime given that a component has survived or not failed until time t. The MRL function is defined by

$$r(t \mid \boldsymbol{\theta}) = E\left[T-t \mid T>t\right] = \frac{1}{S(t \mid \boldsymbol{\theta})}\int_t^\infty y f(y \mid \boldsymbol{\theta})dy - t,$$

where $f(t \mid \boldsymbol{\theta})$ and $S(t \mid \boldsymbol{\theta})$ are, respectively, the PDF and survival function of the random variable T, and $\boldsymbol{\theta}$ is the parameter vector.

Proposition 3.2.1. *The MRL function of the random variable $T \sim RWL(\mu,\phi)$ is given by*

$$r(t \mid \mu,\phi) = \frac{2\mu}{[a(\mu,\phi)+2\mu\phi]\Gamma(\phi)S(t \mid \mu,\phi)}\left[\Gamma\left(\phi+1,\frac{a(\mu,\phi)t}{2\mu}\right) + \frac{2\mu\Gamma\left(\phi+2,\frac{a(\mu,\phi)t}{2\mu}\right)}{a(\mu,\phi)}\right] - t,$$

where $S(t \mid \mu,\phi)$ is the survival function defined in Equation (3.4).

Proof. By using the mixture representation given in Equation (3.2), we have

$$\begin{aligned}\int_t^\infty y f(y \mid \boldsymbol{\theta})dy = \left(\frac{a(\mu,\phi)}{a(\mu,\phi)+2\mu\phi}\right)\int_t^\infty y f_1(y \mid \mu,\phi)dy \\ + \left(\frac{2\mu\phi}{a(\mu,\phi)+2\mu\phi}\right)\int_t^\infty y f_2(y \mid \mu,\phi)dy, \quad (3.7)\end{aligned}$$

where $f_j(y \mid \mu,\phi) = \left(\frac{a(\mu,\phi)}{2\mu}\right)^{\phi+j-1}\frac{y^{\phi+j-2}}{\Gamma(\phi+j-1)}\exp\left\{-\frac{a(\mu,\phi)}{2\mu}y\right\}$, for $j=1,2$.

Now, for $j = 1, 2$,

$$\begin{aligned}\int_t^\infty y f_j(y \mid \mu, \phi) dy &= \frac{[a(\mu,\phi)]^{\phi+j-1}}{(2\mu)^{\phi+j-1}\Gamma(\phi+j-1)} \int_t^\infty y^{\phi+j-1} \exp\left\{-\frac{a(\mu,\phi)}{2\mu} y\right\} dy \\ &= \frac{2\mu}{a(\mu,\phi)\Gamma(\phi+j-1)} \int_{\frac{a(\mu,\phi)t}{2\mu}}^\infty z^{\phi+j-1} \exp\{-z\} dz, \qquad \left(z = \frac{a(\mu,\phi)y}{2\mu}\right) \\ &= \frac{2\mu\Gamma\left(\phi+j, \frac{a(\mu,\phi)t}{2\mu}\right)}{a(\mu,\phi)\Gamma(\phi+j-1)}. \end{aligned} \tag{3.8}$$

Substituting Equation (3.8) into Equation (3.7), we can get the result after some algebraic manipulations. ■

Figure 4 shows the possible shapes for the MRL function of the RWL distribution. Note that as the hazard function is bathtub-shaped (increasing), the MRL function has upside-down bathtub (monotonically decreasing) shape according to Bryson and Siddiqui (1969) and Olcay (1995).

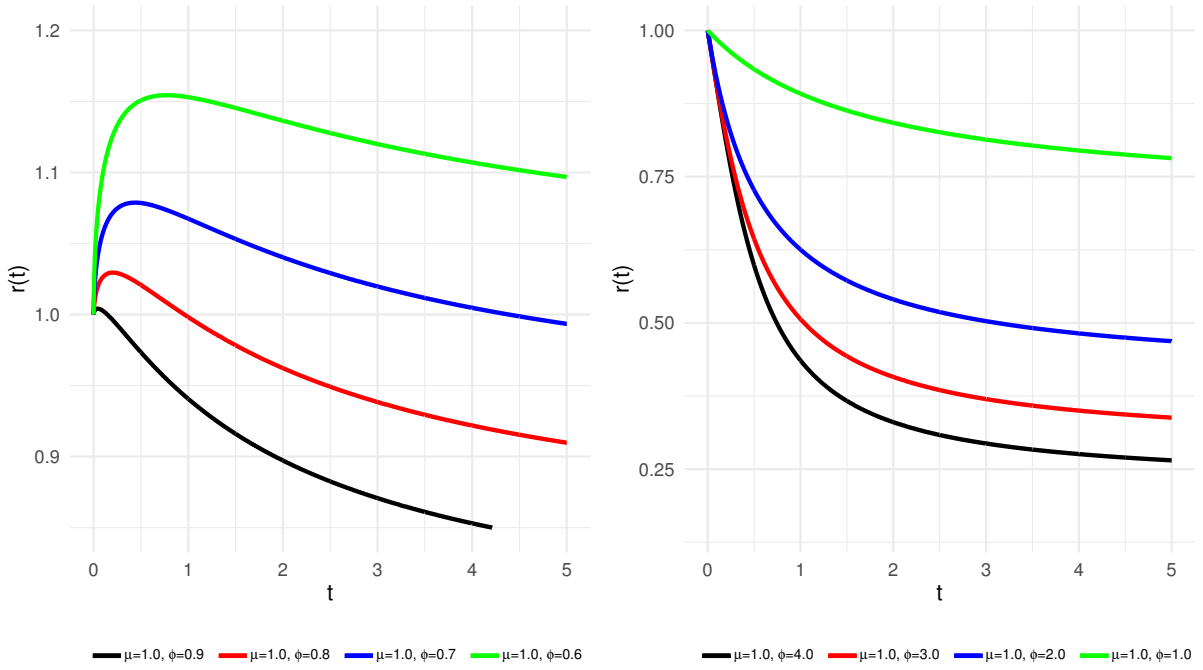

Figure 4 – Plots of the MRL function of the RWL distribution.

Source: Elaborated by the author.

3.2.4 Mean and median deviations

Mean and median deviations are useful for measuring the amount of scattering in a population. They are defined as follows.

Consider a random variable T with PDF $f(t)$ and let μ and m denote, respectively, the mean and median of T, that is, $\mu = \mathbb{E}[T]$ and $m = Median[T]$. Then, the mean and median deviations are defined, respectively, by

$$\delta_1 = \int_0^\infty |t-\mu| f(t)dt \quad \text{and} \quad \delta_2 = \int_0^\infty |t-m| f(t)dt.$$

After some algebraic manipulations, we find the following simplified expressions for δ_1 and δ_2:

$$\delta_1 = 2\left[\mu F(\mu) - \zeta(\mu)\right] \quad \text{and} \quad \delta_2 = m - 2\zeta(m), \tag{3.9}$$

where $F(\cdot)$ is the CDF of T and $\zeta(\cdot)$ is defined as

$$\zeta(s) = \int_0^s t f(t)dt, \quad s > 0.$$

Proposition 3.2.2. *The mean and median deviations for a random variable $T \sim RWL(\mu,\phi)$ are given, respectively, by*

$$\delta_1 = 2\left[\mu F(\mu) - \frac{2\mu}{[a(\mu,\phi)+2\mu\phi]\Gamma(\phi)}\left(\gamma\left(\phi+1, \frac{a(\mu,\phi)}{2}\right) + \frac{2\mu\gamma\left(\phi+2, \frac{a(\mu,\phi)}{2}\right)}{a(\mu,\phi)}\right)\right], \tag{3.10}$$

and

$$\delta_2 = m - \frac{2\mu}{[a(\mu,\phi)+2\mu\phi]\Gamma(\phi)}\left[\gamma\left(\phi+1, \frac{a(\mu,\phi)}{2\mu}m\right) + \frac{2\mu\gamma\left(\phi+2, \frac{a(\mu,\phi)}{2\mu}m\right)}{a(\mu,\phi)}\right], \tag{3.11}$$

where $F(\cdot)$ is the CDF of the RWL distribution given in Equation (3.6) *and*

$$\gamma(b,c) = \int_0^c t^{b-1}e^{-t}dt,$$

is the lower incomplete gamma function.

Proof. It is enough to solve the integral

$$\zeta(s) = \int_0^s t f(t \mid \mu,\phi)dt, \quad s > 0,$$

where $f(\cdot)$ is the PDF given in Equation (3.1). In fact, using the two-component mixture given in Equation (3.2), we have

$$\zeta(s) = \left(\frac{a(\mu,\phi)}{a(\mu,\phi)+2\mu\phi}\right)\int_0^s t f_1(t \mid \mu,\phi)dt + \left(\frac{2\mu\phi}{a(\mu,\phi)+2\mu\phi}\right)\int_0^s t f_2(t \mid \mu,\phi)dt.$$

where $f_j(t \mid \mu,\phi) = \left(\frac{a(\mu,\phi)}{2\mu}\right)^{\phi+j-1}\frac{t^{\phi+j-2}}{\Gamma(\phi+j-1)}\exp\left\{-\frac{a(\mu,\phi)}{2\mu}t\right\}$, for $j = 1,2$.

Let $z = \frac{a(\mu,\phi)}{2\mu}t$, so $dz = \frac{a(\mu,\phi)}{2\mu}dt$. Thus, after some algebraic manipulations, we get

$$\zeta(s) = \frac{2\mu}{[a(\mu,\phi)+2\mu\phi]\Gamma(\phi)}\left(\gamma\left(\phi+1, \frac{a(\mu,\phi)}{2\mu}s\right) + \frac{2\mu\gamma\left(\phi+2, \frac{a(\mu,\phi)}{2\mu}s\right)}{a(\mu,\phi)}\right). \tag{3.12}$$

Now, the results given in Equations (3.10) and (3.11) follow easily by using Equations (3.9) and (3.12). ■

3.2.5 Laplace transform

The Laplace transform of a PDF is useful in several applications of mathematics, engineering and statistics, such as frailty models, machine learning, complex differential equations, signal processing, control systems, among others.

Proposition 3.2.3. *The Laplace transform of the RWL distribution at a complex argument s is given by*

$$\mathscr{L}_f(s) = \left(\frac{1}{a(\mu,\phi)+2\mu\phi}\right)\left(\frac{a(\mu,\phi)}{2\mu s + a(\mu,\phi)}\right)^{\phi+1}\left[a(\mu,\phi)+2\mu(s+\phi)\right].$$

Proof. Let $\boldsymbol{\theta} = (\mu,\phi)$. Then,

$$\begin{aligned}\mathscr{L}_f(s) &= \int_0^\infty e^{-st} f(t \mid \boldsymbol{\theta})dt \\ &= \left(\frac{a(\mu,\phi)}{a(\mu,\phi)+2\mu\phi}\right)\int_0^\infty e^{-st} f_1(t \mid \boldsymbol{\theta})dt + \left(\frac{2\mu\phi}{a(\mu,\phi)+2\mu\phi}\right)\int_0^\infty e^{-st} f_2(t \mid \boldsymbol{\theta})dt, \quad (3.13)\end{aligned}$$

where $f_j(t \mid \mu,\phi) = \left(\frac{a(\mu,\phi)}{2\mu}\right)^{\phi+j-1}\frac{t^{\phi+j-2}}{\Gamma(\phi+j-1)}\exp\left\{-\frac{a(\mu,\phi)}{2\mu}t\right\}$, for $j = 1,2$.

Now, note that

$$\int_0^\infty e^{-st} f_j(t \mid \mu,\phi)dt = \left(\frac{a(\mu,\phi)}{a(\mu,\phi)+2\mu s}\right)^{\phi+j-1}. \quad (3.14)$$

Thus, by substituting Equation (3.14) into Equation (3.13) and making some algebraic manipulations, we obtain

$$\mathscr{L}_f(s) = \left(\frac{1}{a(\mu,\phi)+2\mu\phi}\right)\left(\frac{a(\mu,\phi)}{2\mu s + a(\mu,\phi)}\right)^{\phi+1}\left[a(\mu,\phi)+2\mu(s+\phi)\right].$$

■

3.3 Estimation

We consider the situation where the lifetime is not completely observed and is subject to random right-censoring. The mechanism of random right-censoring is what most occurs in practical problems and it generalizes the Type I and Type II right-censoring mechanisms (COLOSIMO; GIOLO, 2006).

Let C_i denote the censoring time, and T_i be the lifetime of interest for the i-th sampling unit. Suppose that the random variables C_i and T_i are independent. We then observe $t_i = \min(T_i, C_i)$ and $\nu_i = I(T_i \leq C_i)$, where $\nu_i = 1$ if T_i is the observed lifetime and $\nu_i = 0$ if it is the censoring time. From n pairs of times and censoring indicators $(t_1,\nu_2),(t_2,\nu_2),\ldots,(t_n,\nu_n)$, the observed likelihood function for $\boldsymbol{\theta} = (\mu,\phi)^\top$ under non-informative censoring is given by

$$L(\boldsymbol{\theta} \mid \boldsymbol{t}) = \prod_{i=1}^{n}\left[f(t_i \mid \boldsymbol{\theta})\right]^{\nu_i}\left[S(t_i \mid \boldsymbol{\theta})\right]^{1-\nu_i}, \quad (3.15)$$

where $f(t_i \mid \boldsymbol{\theta})$ and $S(t_i \mid \boldsymbol{\theta})$ are the PDF and survival function of the RWL distribution, defined in Equations (3.1) and (3.4), respectively.

Since $h(t_i \mid \boldsymbol{\theta}) = \dfrac{f(t_i \mid \boldsymbol{\theta})}{S(t_i \mid \boldsymbol{\theta})}$, we then have that the likelihood function (3.15) reduces to

$$L(\boldsymbol{\theta} \mid \boldsymbol{t}) = \prod_{i=1}^{n} [h(t_i \mid \boldsymbol{\theta})]^{\nu_i} S(t_i \mid \boldsymbol{\theta}),$$

where $h(t_i \mid \boldsymbol{\theta})$ is the hazard function of the RWL distribution, given in Equation (3.5). Therefore, the log-likelihood function for $\boldsymbol{\theta}$ can be expressed as

$$\begin{aligned}\ell(\boldsymbol{\theta} \mid \boldsymbol{t}) &= d(\phi+1)\log[a(\mu,\phi)] + (\phi-1)\sum_{i=1}^{n}\nu_i\log(t_i) + \sum_{i=1}^{n}\nu_i\log(1+t_i) - \frac{a(\mu,\phi)}{2\mu}\sum_{i=1}^{n}\nu_i t_i \\ &- \sum_{i=1}^{n}\nu_i\log\left[(2\mu)^{\phi-1}[a(\mu,\phi)+2\mu\phi]\,\Gamma\left(\phi, \frac{a(\mu,\phi)}{2\mu}t_i\right) + [a(\mu,\phi)t_i]^{\phi}\exp\left\{-\frac{a(\mu,\phi)}{2\mu}t_i\right\}\right] \\ &- d\log(2\mu) - n\log(\Gamma(\phi)) + \sum_{i=1}^{n}\log\left[\Gamma\left(\phi, \frac{a(\mu,\phi)}{2\mu}t_i\right) + \frac{[a(\mu,\phi)t_i]^{\phi}\exp\left\{-\frac{a(\mu,\phi)}{2\mu}t_i\right\}}{(2\mu)^{\phi-1}[a(\mu,\phi)+2\mu\phi]}\right],\end{aligned} \tag{3.16}$$

where $d < n$ is the observed number of failures and $a(\mu,\phi)$ is defined as previously.

The MLE of parameter vector $\boldsymbol{\theta}$ can be found by maximizing the log-likelihood function given in Equation (3.16). In this work, we used the R function `maxLik`, which is available in the package of the same name to carry out such optimization procedure; see (HENNINGSEN; TOOMET, 2011).

When we have uncensored data, $\nu_i = 1, \forall i$. In this case, the MLE of μ is the sample mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} t_i$, whereas the MLE of ϕ is found by maximizing the following log-likelihood function:

$$\begin{aligned}\ell(\phi \mid \boldsymbol{t}) \propto n(\phi+1)\log[a(\hat{\mu},\phi)] + (\phi-1)\sum_{i=1}^{n}\log(t_i) + \sum_{i=1}^{n}\log(1+t_i) - \frac{a(\hat{\mu},\phi)}{2\hat{\mu}}\sum_{i=1}^{n}t_i \\ - n\phi\log(2\hat{\mu}) - n\log(\Gamma(\phi)) - n\log(a(\hat{\mu},\phi)+2\hat{\mu}\phi),\end{aligned}$$

which can be made by using, for example, the `maxLik` function.

Under mild conditions, it can be shown that the MLE $\hat{\boldsymbol{\theta}}$ is consistent and follows an asymptotic bivariate normal distribution with mean vector $\boldsymbol{\theta}$ and covariance matrix equal to the inverse of the expected Fisher information matrix $\mathscr{I}(\boldsymbol{\theta})$, that is,

$$(\hat{\mu},\hat{\phi}) \xrightarrow{D} \mathrm{N}_2\left((\mu,\phi), \mathscr{I}^{-1}(\mu,\phi)\right), \quad \text{as} \quad n \to \infty,$$

where $\xrightarrow{D}$ denotes convergence in distribution. Unfortunately, the exact expected Fisher information matrix is difficult to be obtained for the RWL distribution. In this case, we can approximate it by its observed version obtained from the `maxLik` package results. Hence, we can construct approximate $100(1-\alpha)\%$ confidence intervals for the individual parameters, along with hypothesis tests, through the estimated marginal distributions (both normal).

3.4 Results based on computation

In this section, we perform a Monte Carlo simulation study to verify the asymptotic behavior of MLEs of the RWL distribution parameters under different sample sizes and percentages of censoring. All the analyses were carried out using the R software (R Core Team, 2021), and the seed used in the pseudo-random number generators was 2020. The random samples of size n from the RWL distribution with parameters μ and ϕ were generated by adapting the Algorithm 1.

To evaluate the MLEs, the following performance criteria were considered: mean relative estimate (MRE) and mean squared error (MSE), which are given, respectively, by

$$\text{MRE}_i = \frac{1}{N}\sum_{j=1}^{N}\frac{\hat{\theta}_{i,j}}{\theta_i} \quad \text{and} \quad \text{MSE}_i = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_{i,j}-\theta_i\right)^2, \quad i=1,2,$$

where $\boldsymbol{\theta} = (\theta_1,\theta_2)^\top = (\mu,\phi)^\top$ is the parameter vector and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1,\hat{\theta}_2)^\top = (\hat{\mu},\hat{\phi})^\top$ is its MLE, while $N = 10{,}000$ is the number of estimates obtained through the proposed approach.

According to these criteria, it is expected that the MRE and MSE return values close to one and zero, respectively. We also compute the coverage probabilities (CPs) of the 95% confidence intervals. For a large number of experiments using 95% confidence intervals (CIs), the relative frequencies of these intervals that covered the true values of $\boldsymbol{\theta}$ should be closer to 0.95. The CPs were calculated using the numeric observed information matrix obtained from the `maxLik` package results.

We considered a sample size $n \in \{20,50,100,200,400\}$ and $\boldsymbol{\theta} \in \{(0.5,0.7),(2,5)\}$, with censoring percentages of 0%, 25% and 50%. We selected these values for $\boldsymbol{\theta}$ in order to get, respectively, bathtub-shaped and increasing hazard functions. The censoring times C_i are generated from the Uniform$(0,\alpha)$ distribution, with the values of $\alpha > 0$ depending on n and of censoring proportion (p). Let p_j, $j = 1,2,\ldots,N$, denote the proportion of censored data in the j-th sample, then according to this procedure it is expected that the mean for the proportions of censored data $(\hat{p})$ will be approximately 0, 0.25 and 0.5.

Under these scenarios, we report the values of the empirical MREs, MSEs, and CPs in Tables 1 and 2. According to these tables, we can see that the MSEs of all estimators tend to zero as the sample size increases, suggesting that all estimators are consistent with the parameters. In contrast, the MRE values tend to one, meaning that the estimators are asymptotically unbiased for the parameters, as expected. We can also see that, as the censoring percentage increases, the MREs and MSEs of the MLEs also increase, as expected. Furthermore, we observe that, as n increases, the CPs tend to the nominal level (0.95). Therefore, in general, all of these results show the excellent performance of the MLEs of the corresponding parameters.

Table 1 – MRE, MSE, CP and expected censoring proportion estimates for $N = 10,000$ samples of sizes $n \in \{20, 50, 100, 200, 400\}$, with 0%, 25% and 50% of random censored data, for $\mu = 0.5$ and $\phi = 0.7$.

		$\mu = 0.5$			$\phi = 0.7$			
	n	MRE	MSE	CP	MRE	MSE	CP	$\hat{p}$
0%	20	1.001	0.015	0.910	1.147	0.080	0.964	-
	50	1.000	0.006	0.932	1.052	0.021	0.950	-
	100	0.999	0.003	0.942	1.025	0.009	0.955	-
	200	1.001	0.002	0.949	1.013	0.004	0.948	-
	400	1.001	0.001	0.944	1.006	0.002	0.950	-
25%	20	1.024	0.027	0.904	1.153	0.098	0.956	0.250
	50	1.011	0.010	0.931	1.052	0.025	0.954	0.250
	100	1.007	0.004	0.946	1.025	0.011	0.952	0.250
	200	1.005	0.002	0.948	1.012	0.005	0.950	0.250
	400	1.002	0.001	0.945	1.007	0.002	0.955	0.249
50%	20	1.093	0.088	0.888	1.192	0.168	0.959	0.500
	50	1.031	0.020	0.919	1.063	0.034	0.957	0.500
	100	1.017	0.009	0.938	1.030	0.014	0.953	0.500
	200	1.010	0.004	0.945	1.015	0.007	0.952	0.500
	400	1.004	0.002	0.946	1.008	0.003	0.951	0.499

Source: Elaborated by the author.

Table 2 – MRE, MSE, CP and expected censoring proportion estimates for $N = 10,000$ samples of sizes $n \in \{20, 50, 100, 200, 400\}$, with 0%, 25% and 50% of random censored data, for $\mu = 2$ and $\phi = 5$.

		$\mu = 2$			$\phi = 5$			
	n	MRE	MSE	CP	MRE	MSE	CP	$\hat{p}$
0%	20	1.002	0.038	0.929	1.150	4.548	0.942	-
	50	1.002	0.015	0.936	1.066	1.457	0.949	-
	100	1.001	0.008	0.943	1.032	0.622	0.952	-
	200	1.001	0.004	0.950	1.017	0.292	0.952	-
	400	1.001	0.002	0.945	1.009	0.141	0.951	-
25%	20	1.004	0.049	0.927	1.180	6.610	0.925	0.249
	50	1.002	0.019	0.943	1.074	1.806	0.953	0.250
	100	1.001	0.009	0.948	1.035	0.791	0.952	0.250
	200	1.001	0.005	0.948	1.019	0.369	0.951	0.250
	400	1.001	0.002	0.946	1.011	0.178	0.947	0.250
50%	20	1.010	0.080	0.927	1.232	11.910	0.902	0.498
	50	1.003	0.029	0.942	1.092	2.621	0.943	0.500
	100	1.002	0.014	0.947	1.045	1.155	0.951	0.501
	200	1.001	0.007	0.950	1.024	0.537	0.949	0.500
	400	1.001	0.003	0.948	1.013	0.252	0.951	0.501

Source: Elaborated by the author.

3.5 Real data examples

In this section, we illustrate the proposed methodology on electrical appliances data (Section 3.5.1), as well as on lifetimes of an agricultural machine (Section 3.5.2).

We compared the results obtained by the RWL distribution with the corresponding ones achieved with the use of other two-parameter lifetime distributions reparameterized by their mean. Namely, the reparameterized inverse gamma, and reparameterized Birnbaum-Saunders distributions (BOURGUIGNON; GALLARDO, 2020; SANTOS-NETO *et al.*, 2012). We present the PDFs of these distributions as follows:

- **Reparameterized inverse gamma (RIG) distribution:**
  According to Bourguignon and Gallardo (2020), the PDF of the RIG distribution is given by

$$f(t \mid \mu,\phi) = \frac{[\mu(1+\phi)]^{\phi+2}}{\Gamma(\phi+2)} t^{-\phi-3} \exp\left\{-\frac{\mu(1+\phi)}{t}\right\}, \quad t>0,$$

  where $\mu > 0$ is the mean parameter and $\phi > 0$ is the precision parameter.

- **Reparameterized Birnbaum-Saunders (RBS) distribution:**
  Presented by Santos-Neto *et al.* (2012), it has PDF given by

$$f(t \mid \mu,\phi) = \frac{\exp\{\phi/2\}\sqrt{\phi+1}}{4t^{3/2}\sqrt{\pi\mu}} \left(t+\frac{\phi\mu}{\phi+1}\right) \exp\left\{-\frac{\phi}{4}\left[\frac{(\phi+1)t}{\phi\mu}+\frac{\phi\mu}{(\phi+1)t}\right]\right\},$$

  for all $t > 0$, where $\mu > 0$ is the mean parameter and $\phi > 0$ is the precision parameter.

In order to carry out the model selection, different discrimination criterion methods based on log-likelihood function evaluated at the MLEs were considered. Let k be the number of parameters in the model and $\hat{\boldsymbol{\theta}}$ denote the MLE for the parameter vector $\boldsymbol{\theta}$. Then, the model discrimination criteria used here are: Akaike Information Criterion (AIC) (AKAIKE, 1974), Corrected AIC (AICc) (SUGIURA, 1978), Bayesian or Schwarz Information Criterion (BIC) (SCHWARZ, 1978), Hannan-Quinn Information Criterion (HQIC) (HANNAN; QUINN, 1979), and Consistent AIC (CAIC) (BOZDOGAN, 1987), which are computed, respectively, by

$$\begin{aligned}
\text{AIC} &= -2\ell\left(\hat{\boldsymbol{\theta}};\boldsymbol{t}\right)+2k,\\
\text{AICc} &= \text{AIC}+\frac{2k(k+1)}{(n-k-1)},\\
\text{BIC} &= -2\ell\left(\hat{\boldsymbol{\theta}};\boldsymbol{t}\right)+k\log(n),\\
\text{HQIC} &= -2\ell\left(\hat{\boldsymbol{\theta}};\boldsymbol{t}\right)+2k\log\left(\log(n)\right),\\
\text{CAIC} &= \text{AIC}+k\left[\log(n)-1\right],
\end{aligned}$$

where $\ell(.\mid \boldsymbol{t})$ is the log-likelihood function of the corresponding model and n is the sample size. According to these criteria, the best model is the one that provides the minimum values. The

Kolmogorov-Smirnov test with confidence level $\alpha = 0.05$ was also considered for checking the goodness-of-fit of models to the uncensored data (DANIEL, 1990).

3.5.1 Cycles up to the failure for electrical appliances

In this subsection, we reanalyzed the data set extracted from Lawless (2011), which consists of a number of cycles, divided by 1,000, up to the failure for 60 electrical appliances in a life test (see Table 3). Many authors have analyzed these uncensored data, including Reed (2011), Khan (2018) and Ramos *et al.* (2019). Such data are known to have a bathtub-shaped hazard function.

Table 3 – Number of cycles, divided by 1,000, up to the failure for 60 electrical appliances in a life test.

0.014	0.034	0.059	0.061	0.069	0.080	0.123	0.142	0.165	0.210
0.381	0.464	0.479	0.556	0.574	0.839	0.917	0.969	0.991	1.064
1.088	1.091	1.174	1.270	1.275	1.355	1.397	1.477	1.578	1.649
1.702	1.893	1.932	2.001	2.161	2.292	2.326	2.337	2.628	2.785
2.811	2.886	2.993	3.122	3.248	3.715	3.790	3.857	3.912	4.100
4.106	4.116	4.315	4.510	4.580	5.267	5.299	5.583	6.065	9.701

Source: Lawless (2011).

Table 4 displays the MLEs, standard errors (SEs) and 95% confidence intervals (CIs) for the parameters μ and ϕ of the RWL model. Note that the estimated mean number of cycles to failure of an electrical appliance is 2.193 cycles. Furthermore, since $\hat{\phi} = 0.733$, the estimated hazard function is bathtub-shaped, that is, it is characterized by an increased number of failures (and thus, unavailability) in the initial period of electrical appliance usage after its commissioning, followed by a long span of normal use with a small and roughly constant number of failures, and finally, a period of a fast increasing number of failures occurring because of the age of the observed electrical appliance.

Table 4 – MLEs, SEs and 95% CIs for the parameters of the RWL distribution, considering the electrical appliances data.

Parameter	MLE	SE	95% CI
μ	2.193	0.272	[1.659; 2.727]
ϕ	0.733	0.136	[0.466; 1.001]

Source: Elaborated by the author.

Table 5 gives the log-likelihood, AIC, AICc, BIC, HQIC, and CAIC values, along with the Kolmogorov-Smirnov (KS) test statistics and their p-values, for all four distributions considered. We can see that the RWL distribution offers a better fit to the electrical appliances data than the other models considered since it has the minimum values of these criteria. In

addition, the KS test indicates that the electrical appliances data are a random sample from a RWL distribution with $\hat{\mu} = 2.193$ and $\hat{\phi} = 0.733$.

Table 5 – Model selection criteria values and KS test (statistic and p-values) for the fitted probability distributions, considering the electrical appliances data.

Criterion	RWL	RIG	RBS
Log-likelihood	-105.774	-157.273	-118.912
AIC	215.548	318.546	241.824
AICc	215.759	318.756	242.035
BIC	219.737	322.734	246.013
HQIC	217.187	320.184	243.463
CAIC	221.737	324.734	248.013
KS statistic	0.072	0.496	0.285
p-value	0.907	< 0.0001	< 0.001

Source: Elaborated by the author.

Figure 5 presents the survival function adjusted by different probability distributions (RWL, RIG and RBS distributions) superimposed to the estimated Kaplan-Meier (KM) survival curve. From this figure, it can be observed that the RWL distribution provides the best fit to the electrical appliances data. Therefore, from the proposed methodology, the data set related to the failure times of 60 electrical appliances can be well-described by the RWL distribution.

3.5.2 Agricultural machine data

As a second application, in this subsection we reanalyzed the data related to the times up to corrective maintenance of an agricultural machine, presented by Ramos *et al.* (2019). This data set includes two censored observations, both in 13 days. Its analysis can be useful to correctly predict the next maintenance in order to reduce costs.

Table 6 – Times up to corrective maintenance of an agricultural machine ("+" denotes censoring).

1	1	1	1	1	1	1	2	2	3	3	3
3	3	4	4	4	4	4	4	4	5	5	5
5	5	5	5	5	5	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7
7	7	8	8	8	8	8	8	8	8	8	8
8	9	9	9	9	9	11	11	11	11	11	11
11	11	13	13+	13+	-	-	-	-	-	-	-

Source: Ramos *et al.* (2019).

Table 7 shows the MLEs, SEs and 95% CIs for the parameters μ and ϕ of the RWL distribution. Notice that the estimated mean time to occur a fail in the agricultural machine is

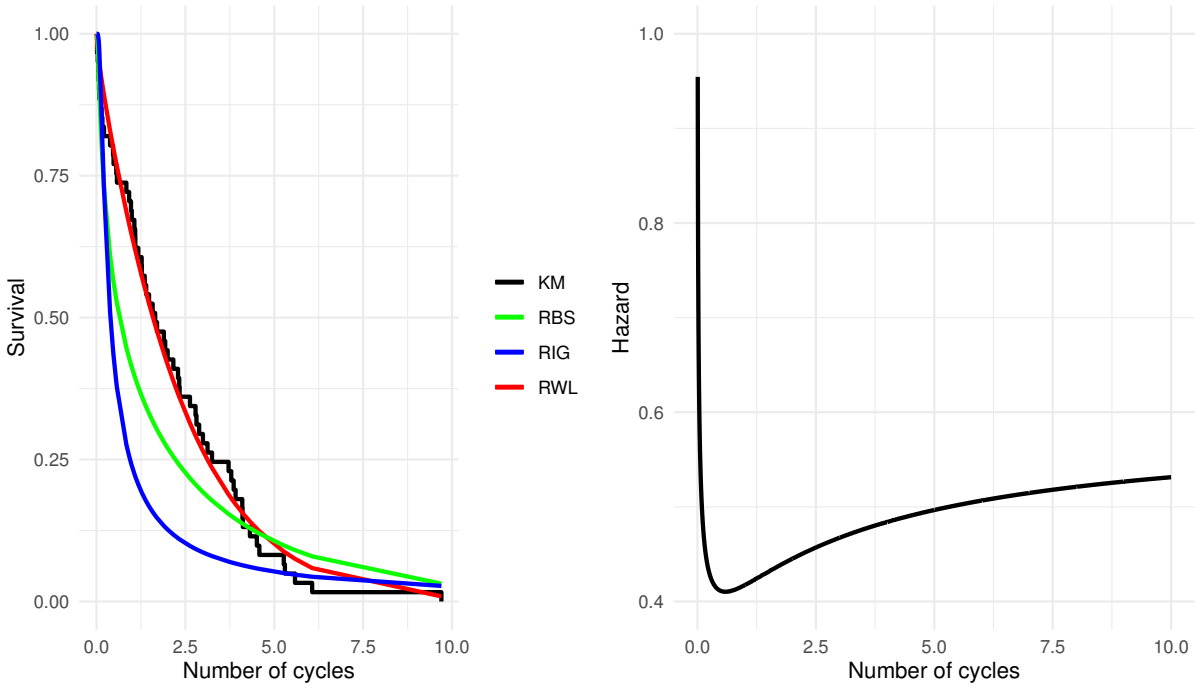

Figure 5 – Left panel: Fitted survival functions superimposed to the estimated KM survival curve, considering the electrical appliances data. Right panel: Estimated hazard function of the RWL distribution for these data.

Source: Elaborated by the author.

6.404 days. Also, the fit of the RWL distribution suggests an increasing-shaped hazard function: $\hat{\phi} = 2.778$ (see Figure 6, right panel).

Table 7 – MLEs, SEs and 95% CIs for the parameters of the RWL distribution, considering the agricultural machine data.

Parameter	MLE	SE	95% CI
μ	6.404	0.369	[5.680; 7.127]
ϕ	2.778	0.491	[1.816; 3.740]

Source: Elaborated by the author.

Table 8 reports the results from different model discrimination/selection criteria, such as the log-likelihood, AIC, AICc, BIC, HQIC and CAIC, for the four considered probability distributions. From these results, we see that the RWL distribution provides slightly better description of the data compared to other candidate distributions, since it yields the lowest values in all criteria.

Table 8 – Model selection criteria values for the fitted probability distributions, considering the agricultural machine data.

Criterion	RWL	RIG	RBS
Log-likelihood	−223.049	−248.159	−235.404
AIC	450.098	500.318	474.808
AICc	450.237	500.457	474.947
BIC	455.075	505.295	479.785
HQIC	452.104	502.324	476.814
CAIC	457.075	507.295	481.785

Source: Elaborated by the author.

Figure 6 exhibits the survival functions superimposed to the estimated KM survival curve (left panel), as well as the estimated hazard function (right panel). From this figure, it can be observed that the RWL distribution provides a good fit to the agricultural machine data. Therefore, from the proposed methodology, the data set related to the failure times of agricultural machine can be well-described by the RWL distribution.

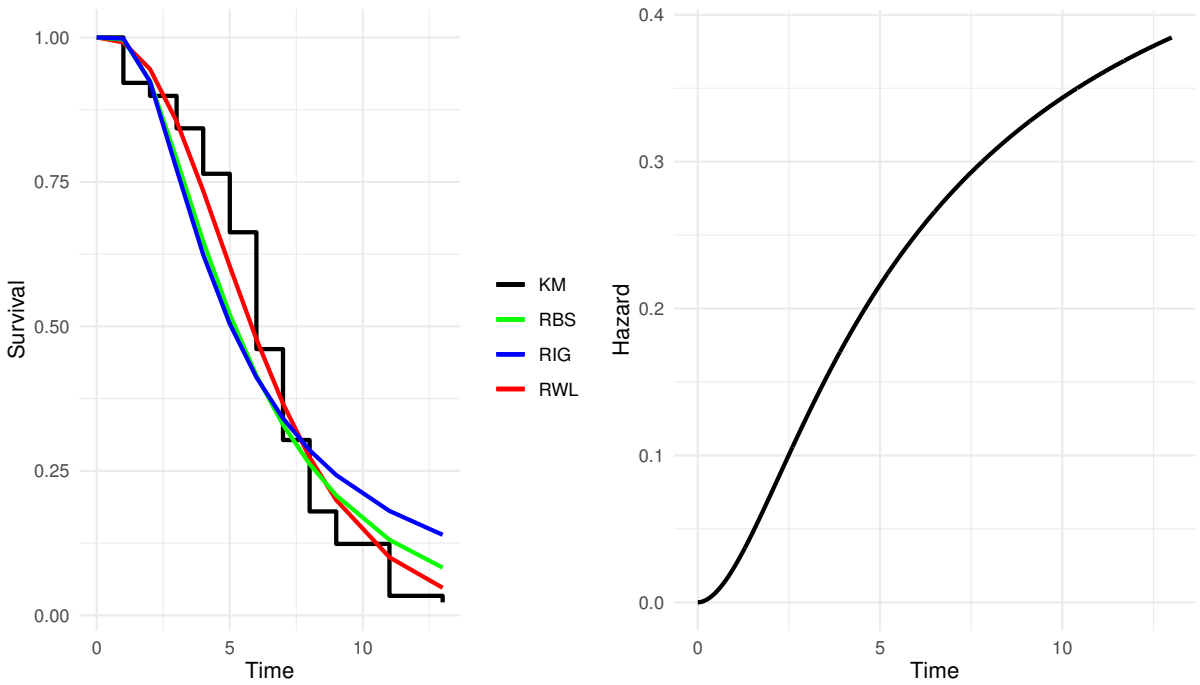

Figure 6 – Left panel: Fitted survival functions superimposed to the estimated KM survival function, considering the lifetimes of an agricultural machine. Right panel: Estimated hazard function of the RWL distribution for these data.

Source: Elaborated by the author.

A preventive approach for this agricultural machine is given as follows. Through the quantile function of the RWL distribution given in Equation (3.6), we can get the number of days that are expected to have a certain percentage of failures. Table 9 displays different times to

failure, assuming different percentages. The results obtained from this table show that preventive maintenance can be performed assuming different percentages of failures. Accordingly, we recommend the agricultural enterprise to consider 4 days (25% of failures) after the last failure to perform maintenance in this agricultural machine.

Table 9 – Days to perform preventive maintenance to agricultural machine by assuming different percentages of failures, based on the RWL distribution.

10%	25%	50%	75%	99%
2.55	3.88	5.82	8.29	16.87

Source: Elaborated by the author.

3.6 Concluding remarks

In this chapter, we derived critical mathematical properties of the RWL distribution such as power, logarithmic and inverse moments, coefficient of variation, harmonic mean, mean and median deviations, Laplace transform among others. Under this parameterization, one of the parameters is given by the mean, whereas the other parameter can be interpreted as a precision parameter. The ML method for the parameters was discussed under random right-censoring. An extensive Monte Carlo simulation study showed that the proposed estimators are consistent and return reasonable estimates for the parameters of the RWL distribution. The proposed methodology was used in two applications considering electrical appliances data and a data set related to the lifetimes of an agricultural machine, in which we observed that the RWL distribution returned best fit when compared to some well-known reparameterized models in the statistical literature.

CHAPTER

4

A WEIGHTED LINDLEY FRAILTY MODEL: ESTIMATION AND APPLICATION TO A LUNG CANCER DATASET

Traditional survival models implicitly assume a homogeneous population to be studied (like in Chapter 3), but covariate information can be included to explain the observable heterogeneity. However, a portion of unobserved heterogeneity can be induced by several factors, such as environmental or genetic factors, or information that was not considered in planning study. Failing to account for this latter form of heterogeneity may lead to distorted results. In this chapter, we introduce a novel frailty model for modeling unobserved heterogeneity in univariate survival data. In this model, the RWL distribution is employed to describe the unobserved frailty. For our model to be identifiable, we use the RWL distribution with mean one as frailty distribution. By using the Laplace transform function of the RWL distribution, both unconditional survival and hazard functions are obtained. We assume Weibull and Gompertz distributions as the baseline hazard functions. Classical inference based on the ML method is developed. Simulation studies are performed to verify the behavior of MLEs under different proportions of right-censoring and to check the likelihood ratio (LR) test for detecting unobserved heterogeneity under different sample sizes. Finally, to demonstrate the applicability of the proposed model, we use it to analyze a medical dataset from a population-based study of incident cases of lung cancer diagnosed in the state of São Paulo, Brazil.

4.1 RWL frailty model

In conditional model (2.23), we suppose that the frailty variable $Z \sim \text{RWL}(1, \phi)$. Then, according to Equation (3.1), the frailty PDF is given by

$$f(z \mid \phi) = \frac{\left(\sqrt{\phi(\phi+1)}\right)^{\phi+1} z^{\phi-1}(1+z)\exp\left(-z\sqrt{\phi(\phi+1)}\right)}{\left(\sqrt{\phi(\phi+1)}+\phi\right)\Gamma(\phi)}, \quad z>0, \tag{4.1}$$

where ϕ is the (unknown) shape parameter. As mentioted in Chapter 2, problems of identifiability in the resulting model can occur. To avoid this, we have assumed $\mathbb{E}[Z] = 1$; see (ELBERS; RIDDER, 1982).

Usually, in a frailty model, the amount of unobserved heterogeneity in a study population is quantified by the variance of the frailty variable. If we assume the PDF from Equation (4.1) to be a frailty distribution, the variance is $\theta = 2\left(\phi + \sqrt{\phi(\phi+1)}\right)^{-1}$. Notice that this variance decreases as ϕ increases, and it tends to infinite when ϕ tends to zero. Thus, small values of ϕ indicate the presence of higher unobserved heterogeneity among subjects.

From Proposition 3.2.3, we have that the Laplace transform of the frailty PDF (4.1) depending on its variance, θ, is given by

$$\mathscr{L}_f(s) = \left(1+\frac{s\theta(\theta+4)}{2(\theta+2)}\right)^{-\frac{4}{\theta(\theta+4)}-1}\left(1+\frac{s\theta}{2}\right), \quad s\in\mathbb{R}. \tag{4.2}$$

If we evaluate Equation (4.2) at $s = H_0(t)\xi$, where $\xi = \exp(\boldsymbol{x}^\top\boldsymbol{\beta})$ for the sake of simplicity, we find that unconditional survival function (2.25) with RWL-distributed frailty, is given by

$$S(t \mid \boldsymbol{x}) = \left(1+\frac{H_0(t)\xi\theta(\theta+4)}{2(\theta+2)}\right)^{-\frac{4}{\theta(\theta+4)}-1}\left(1+\frac{H_0(t)\xi\theta}{2}\right), \quad t>0. \tag{4.3}$$

Then, the corresponding unconditional hazard function (2.26) becomes

$$h(t \mid \boldsymbol{x}) = h_0(t)\xi\left(\frac{4+\theta(\theta+4)}{2(\theta+2)+H_0(t)\xi\theta(\theta+4)} - \frac{\theta}{2+H_0(t)\xi\theta}\right), \quad t>0. \tag{4.4}$$

Hereafter, we will call the model with unconditional survival and hazard functions as RWL frailty model. As mentioned in Chapter 2, the baseline hazard function can be assumed parametrically or non-parametrically. In this work, we use the parametric approach by employing Weibull and Gompertz models as baseline hazard functions. These models are indexed by two parameters and can accommodate constant, increasing, and decreasing hazard functions. The Weibull distribution is often used in survival and reliability, whereas the Gompertz distribution is preferred in actuarial and demographic applications (WIENKE, 2010; BÖHNSTEDT; GAMPE; PUTTER, 2021). However, an advantage of using the Gompertz distribution as baseline hazard

function in survival settings is that sometimes it may be “defective" (ROCHA *et al.*, 2016; CALSAVARA *et al.*, 2019b; CALSAVARA *et al.*, 2019c). This implies that not all subjects are susceptible to the event of interest in the study, that is, they are considered cured or long-term survivors; see (MALLER; ZHOU, 1996).

4.1.1 RWL frailty model with Weibull baseline hazard function

The baseline hazard and cumulative hazard functions of the Weibull distribution are given, respectively, by

$$h_0(t) = \frac{\kappa t^{\kappa-1}}{\rho^{\kappa}} \quad \text{and} \quad H_0(t) = \left(\frac{t}{\rho}\right)^{\kappa}, \quad t > 0, \tag{4.5}$$

where $\kappa > 0$ is the shape parameter and $\rho > 0$ is the scale parameter. For $\kappa < 1$, the Weibull hazard function decreases monotonously; when $\kappa = 1$ (exponential distribution), it is constant over time; and when $\kappa > 1$, this function increases monotonously (WIENKE, 2010).

Therefore, using (4.5) into (4.3) and (4.4), the unconditional survival and hazard functions of the RWL frailty model with Weibull baseline hazard function are, respectively,

$$S(t \mid \boldsymbol{x}) = \left(1 + \frac{t^{\kappa}\xi\theta(\theta+4)}{2\rho^{\kappa}(\theta+2)}\right)^{-\frac{4}{\theta(\theta+4)}-1} \left(1 + \frac{t^{\kappa}\xi\theta}{2\rho^{\kappa}}\right), \quad t > 0 \tag{4.6}$$

and

$$h(t \mid \boldsymbol{x}) = \kappa t^{\kappa-1}\xi \left(\frac{4+\theta(\theta+4)}{2\rho^{\kappa}(\theta+2)+t^{\kappa}\xi\theta(\theta+4)} - \frac{\theta}{2\rho^{\kappa}+t^{\kappa}\xi\theta}\right), \quad t > 0 \tag{4.7}$$

Figure 7 presents some examples of the shapes obtained for survival and hazard functions of the RWL frailty model with Weibull baseline hazard function when selected values of the parameters were used. With this graphical analysis, it is observed that the survival function of this model is always proper, that is, $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(t) = 0$. In addition, the hazard function presents monotonically decreasing and unimodal shapes. It is important to point out that the unimodal shape is not possible for the hazard function of the traditional Weibull model.

4.1.2 RWL frailty model with Gompertz baseline hazard function

The baseline hazard and cumulative hazard functions of the Gompertz distribution are, respectively, given by

$$h_0(t) = \rho e^{\kappa t} \quad \text{and} \quad H_0(t) = \frac{\rho}{\kappa}\left(e^{\kappa t} - 1\right), \quad t > 0, \tag{4.8}$$

where $\kappa > 0$ and $\rho > 0$ are shape and scale parameters, respectively. If $\kappa < 0$, the Gompertz distribution is “defective" (ROCHA *et al.*, 2016; CALSAVARA *et al.*, 2019b; CALSAVARA *et al.*, 2019c), since its cumulative hazard function converges to the constant $-\frac{\rho}{\kappa}$ for $t \to \infty$, which leads to a cure or long-term survivors fraction $p_0 = \exp\left(\frac{\rho}{\kappa}\right)$ in the study population. When

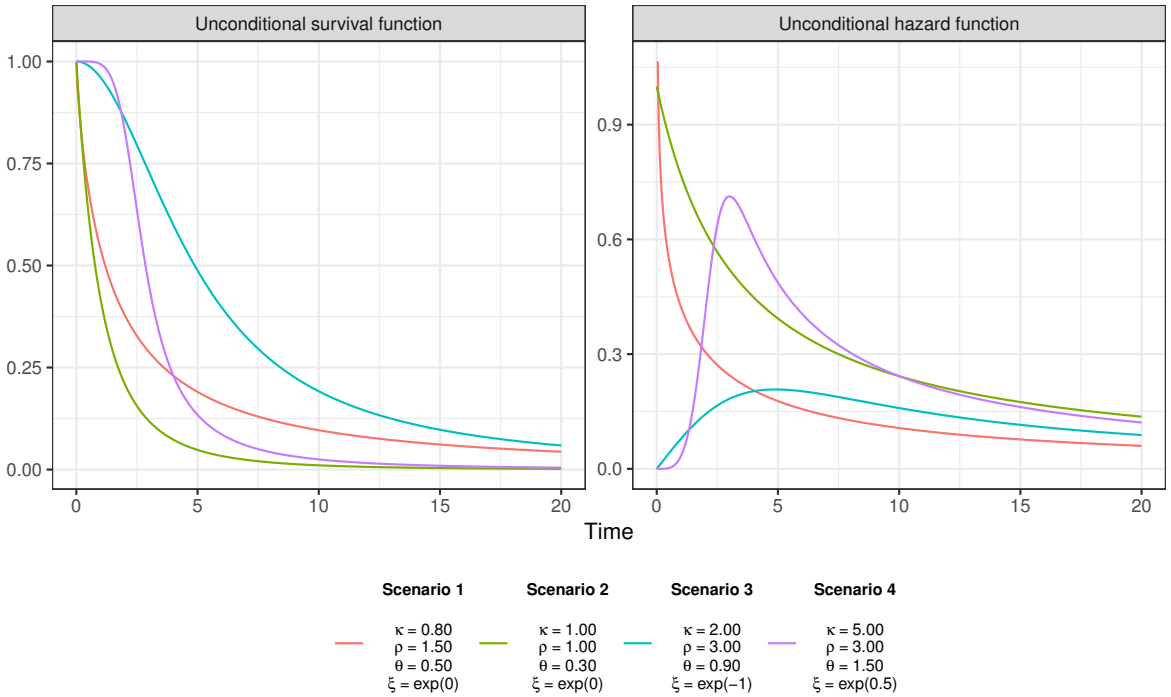

Figure 7 – Unconditional survival (left panel) and hazard (right panel) functions of the RWL frailty model with Weibull baseline hazard function.

Source: Elaborated by the author.

$\kappa = 0$, the exponential distribution is obtained as a special case. Thus, the hazard function of the Gompertz distribution can be decreasing ($\kappa < 0$), constant ($\kappa = 0$) or increasing ($\kappa > 0$).

Using (4.8) into (4.3) and (4.4), the marginal survival and hazard functions of the RWL frailty model with Gompertz baseline hazard function are, respectively,

$$S(t \mid \boldsymbol{x}) = \left(1 + \frac{\rho(e^{\kappa t} - 1)\xi\theta(\theta+4)}{2\kappa(\theta+2)}\right)^{-\frac{4}{\theta(\theta+4)}-1} \left(1 + \frac{\rho(e^{\kappa t} - 1)\xi\theta}{2\kappa}\right), \tag{4.9}$$

and

$$h(t \mid \boldsymbol{x}) = \rho\kappa e^{\kappa t}\xi \left(\frac{4 + \theta(\theta+4)}{2\kappa(\theta+2) + \rho(e^{\kappa t} - 1)\xi\theta(\theta+4)} - \frac{\theta}{2\kappa + \rho(e^{\kappa t} - 1)\xi\theta}\right), \tag{4.10}$$

for all $t > 0$.

If $\kappa > 0$, the marginal survival function (4.9) is proper, that is, $\lim_{t\to 0} S(t \mid \boldsymbol{x}) = 1$ and $\lim_{t\to\infty} S(t \mid \boldsymbol{x}) = 0$. On the other hand, if $\kappa < 0$, it is an improper marginal survival function, since

$$\lim_{t\to\infty} S(t \mid \boldsymbol{x}) = p_0 = \left(1 - \frac{\rho\xi\theta(\theta+4)}{2\kappa(\theta+2)}\right)^{-\frac{4}{\theta(\theta+4)}-1} \left(1 - \frac{\rho\xi\theta}{2\kappa}\right) \in (0,1), \tag{4.11}$$

where p_0 denotes the cured or long-term survivors fraction in the study population.

Figure 8 displays some examples of the shapes obtained for marginal survival and hazard functions of the RWL frailty model with Gompertz baseline hazard function for selected values

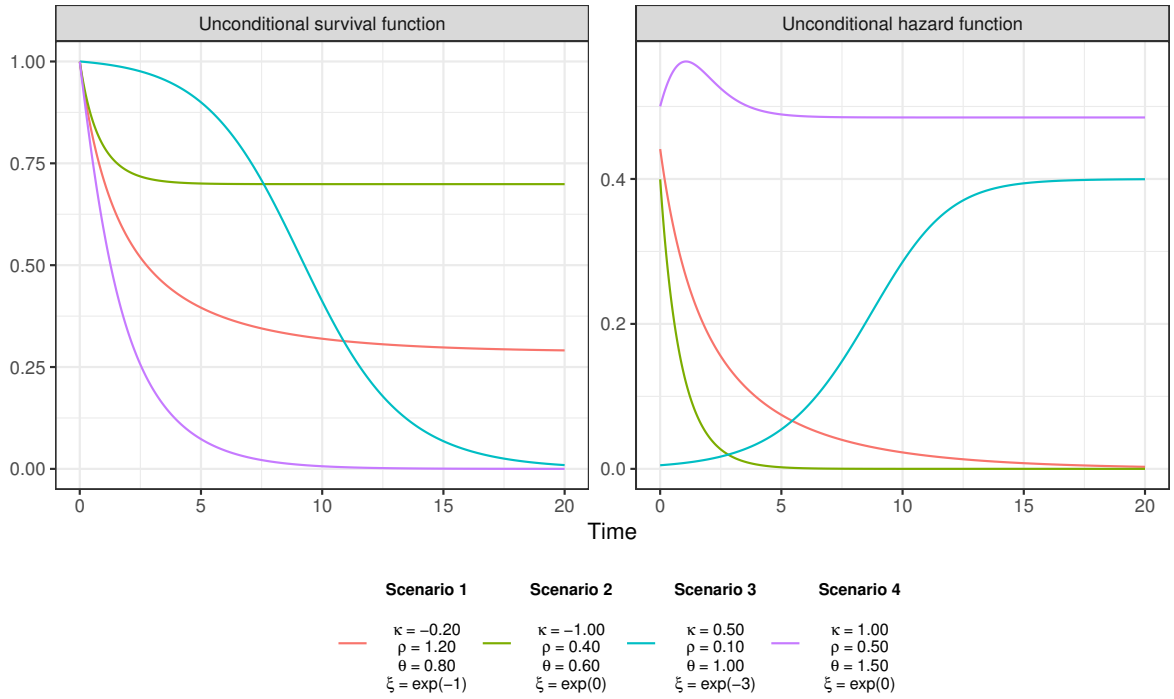

Figure 8 – Unconditional survival (left panel) and hazard (right panel) functions of the RWL frailty model with Gompertz baseline hazard function for some parameter values.

Source: Elaborated by the author.

of the parameters. We can see that the marginal survival function is proper for $\kappa > 0$ and improper for $\kappa < 0$, as previously mentioned. In addition, this model supports unimodal-shaped, monotonically increasing and monotonically decreasing marginal hazard functions. Therefore, the RWL frailty model with Gompertz baseline hazard function is more flexible than the RWL frailty model with Weibull baseline hazard function.

4.2 Inference methods

In this section, we describe the ML method for estimating parameters of the two RWL frailty models. Under certain regularity conditions, MLEs have attractive properties, such as, consistency, efficiency, asymptotic normality, among others (LEHMANN; CASELLA, 2006).

It is possible that lifetime data may not be available for all subjects in a study. For example, some lifetimes are right-censored and it is only known that they are greater than the recorded value. If so, let T_i and C_i be the lifetime and censoring time variables, respectively, for the i^{th} subject in the population under study. Suppose T_i and C_i are independent random variables and let $\nu_i = \mathbb{I}(T_i \leq C_i)$ be the censoring indicator (i.e, $\nu_i = 1$ if T_i is lifetime and $\nu_i = 0$ otherwise). We then observe $t_i = \min(T_i, C_i)$. Let $\boldsymbol{x}_i$ denote a $p \times 1$ vector of covariates, which are observed in the i^{th} subject. Then, from a sample of n subjects, the likelihood function for the

parameter vector $\boldsymbol{\Psi} = (\kappa, \rho, \theta, \boldsymbol{\beta}^\top)^\top$ under non-informative censoring setting is given by

$$\mathscr{L}(\boldsymbol{\Psi}) = \prod_{i=1}^{n} h(t_i|\boldsymbol{x}_i)^{\nu_i} S(t_i|\boldsymbol{x}_i),$$

where $S(\cdot|\boldsymbol{x}_i)$ and $h(\cdot|\boldsymbol{x}_i)$ are unconditional survival and hazard functions for subject i given in Equations (4.3) and (4.4), respectively.

Assuming the Weibull baseline hazard function and using (4.6) and (4.7), the log-likelihood function for $\boldsymbol{\Psi} = (\kappa, \rho, \theta, \boldsymbol{\beta}^\top)^\top$ is expressed as

$$\begin{aligned}\ell(\boldsymbol{\Psi}) = r\log(\kappa) + \sum_{i=1}^{n} \nu_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + (\kappa-1)\sum_{i=1}^{n} \nu_i \log(t_i) + \sum_{i=1}^{n} \log\left(1 + \frac{t_i^\kappa \xi_i \theta}{2\rho^\kappa}\right) \\ - \left(\frac{4}{\theta(\theta+4)} + 1\right) \sum_{i=1}^{n} \log\left(1 + \frac{t_i^\kappa \xi_i \theta(\theta+4)}{2\rho^\kappa(\theta+2)}\right) + \sum_{i=1}^{n} \nu_i \log(\eta_i), \quad (4.12)\end{aligned}$$

where $\eta_i = \frac{4+\theta(\theta+4)}{2\rho^\kappa(\theta+2)+t_i^\kappa\xi_i\theta(\theta+4)} - \frac{\theta}{2\rho^\kappa+t_i^\kappa\xi_i\theta}$, $r = \sum_{i=1}^{n} \nu_i$ is the failure number, and $\xi_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)$.

Now, considering the Gompertz baseline hazard function and using (4.9) and (4.10), the log-likelihood function for $\boldsymbol{\Psi} = (\kappa, \rho, \theta, \boldsymbol{\beta}^\top)^\top$ becomes

$$\begin{aligned}\ell(\boldsymbol{\Psi}) = r\log(\rho\kappa) + \kappa\sum_{i=1}^{n} \nu_i t_i + \sum_{i=1}^{n} \nu_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + \sum_{i=1}^{n} \log\left(1 + \frac{\rho(e^{\kappa t_i}-1)\xi_i\theta}{2\kappa}\right) \\ - \left(\frac{4}{\theta(\theta+4)} + 1\right) \sum_{i=1}^{n} \log\left(1 + \frac{\rho(e^{\kappa t_i}-1)\xi_i\theta(\theta+4)}{2\kappa(\theta+2)}\right) + \sum_{i=1}^{n} \nu_i \log(\vartheta_i), \quad (4.13)\end{aligned}$$

where r and ξ_i are defined as previously, whereas

$$\vartheta_i = \frac{4+\theta(\theta+4)}{2\kappa(\theta+2)+\rho(e^{\kappa t_i}-1)\xi_i\theta(\theta+4)} - \frac{\theta}{2\kappa+\rho(e^{\kappa t_i}-1)\xi_i\theta}.$$

The MLEs $\widehat{\boldsymbol{\Psi}}$ of respective parameter vectors $\boldsymbol{\Psi}$ can be obtained by maximizing the log-likelihood functions (4.12) and (4.13). We note that for two RWL frailty models $\widehat{\boldsymbol{\Psi}}$ does not have a closed-form. Hence, numerical nonlinear optimization methods such as best-performing Broyden-Fletcher-Goldfarb-Shanno (BFGS) or quasi-Newton, Fisher scoring, general-purpose unconstrained nonlinear optimization (cminf) and Simulated ANNealing (SANN) are required in order to find a solution; see (NOCEDAL; WRIGHT, 1999; NIELSEN; MORTENSEN, 2016). These and other optimization methods are implemented in various packages within `R` software (R Core Team, 2021), such as `stats` (VENABLES; RIPLEY, 2013), `maxLik` (HENNINGSEN; TOOMET, 2011), and `optimx` (NASH *et al.*, 2020) packages.

Under standard regularity conditions, fulfilled for parameters in the interior of the parameter space, but not on the boundary, the ML estimator $\widehat{\boldsymbol{\Psi}}$ is consistent and follows a $(p+3)$-variate asymptotic normal distribution with mean $\boldsymbol{\Psi}$ and a $(p+3)\times(p+3)$ variance-covariance

matrix being equal to the inverse of the Fisher information matrix (LEHMANN, 2004). Then, mathematically we have

$$\widehat{\boldsymbol{\Psi}} \xrightarrow{D} \mathcal{N}_{(p+3)}\left(\boldsymbol{\Psi}, \boldsymbol{I}^{-1}(\boldsymbol{\Psi})\right), \quad \text{as } n \to \infty, \tag{4.14}$$

where $\xrightarrow{D}$ represents convergence in distribution and

$$\boldsymbol{I}(\boldsymbol{\Psi}) = \left\{ -\mathbb{E}\left(\frac{\partial^2 \ell(\boldsymbol{\Psi})}{\partial \psi_i \partial \psi_j} \right) \right\},$$

is the $(p+3) \times (p+3)$ Fisher information matrix.

Unfortunately, the Fisher information matrix elements for the two RWL frailty models cannot be obtained analytically, and hence, the Fisher information matrix does not have a mathematical expression form. In this case, we approximate the Fisher information matrix by its observed version, denoted by $\boldsymbol{H}(\widehat{\boldsymbol{\Psi}})$, which is calculated by dropping the expectation operator, $\mathbb{E}(\cdot)$, in the Fisher information matrix, and can be computed numerically without much trouble. Thus, we can rewrite Equation (4.14) as

$$\widehat{\boldsymbol{\Psi}} \xrightarrow{D} \mathcal{N}_{(p+3)}\left(\boldsymbol{\Psi}, \boldsymbol{H}^{-1}(\widehat{\boldsymbol{\Psi}})\right), \quad \text{as } n \to \infty.$$

Therefore, SEs of the MLEs can be computed by taking the square root of the elements diagonal of $\boldsymbol{H}^{-1}(\widehat{\boldsymbol{\Psi}})$. Hence, approximate CIs can subsequently be calculated, and hypothesis tests for the model parameters based on asymptotic marginal normal distributions of the MLEs can be generated. In the next section, we describe a simulation study performed to determine whether the usual asymptotic properties of the MLEs hold and to assess the performance of the LR test to detect unobserved heterogeneity.

4.3 Simulation study

In this section, we report two extensive Monte Carlo simulations. The first simulation evaluated the performance of the MLEs of the two RWL frailty models parameters in response to different sample sizes and censoring proportions. The second simulation was intended to verify the behavior of the LR test to detect unobserved heterogeneity. To assess the effect of covariates on the survival function, we divided the sample into two groups (X). We introduced one regression parameter for $\xi : \xi_i = \exp(\beta_1 x_i)$, where β_1 is the parameter associated to covariate. To introduce random right-censoring, the distribution of censoring times is assumed to be Uniform$(0, \varepsilon)$ where $\varepsilon > 0$ is set to control the proportion of right-censored observations. In all of the simulation studies performed, the datasets (t_i, ν_i, x_i) from the two RWL frailty models were generated by using the Algorithm 2 (for Weibull baseline hazard function) and the Algorithm 3 (for Gompertz baseline hazard function) given as follows

Algorithm 2 – Generator of random numbers from the RWL frailty model with Weibull baseline hazard function

1: Define the values of $\boldsymbol{\Psi} = (\kappa, \rho, \theta, \boldsymbol{\beta}^\top)^\top$.
2: Generate $u \sim \text{Uniform}(0,1)$, then using a numerical method (e.g., `uniroot` function into `stats` package) obtain the lifetime y from the equation:

$$\left(1+\frac{y^{\kappa}\xi\theta(\theta+4)}{2\rho^{\kappa}(\theta+2)}\right)^{-\frac{4}{\theta(\theta+4)}-1}\left(1+\frac{y^{\kappa}\xi\theta}{2\rho^{\kappa}}\right)=u.$$

3: Draw $c \sim \text{Uniform}(0,\varepsilon)$, and compute $t = \min(y,c)$.
4: If $t = y$, then $\nu = 1$, otherwise $\nu = 0$.
5: Repeat the previous steps to obtain the desired sample size.

Algorithm 3 – Generator of random numbers from the RWL frailty model with Gompertz baseline hazard function

1: Define the values of $\boldsymbol{\Psi} = (\kappa, \rho, \theta, \boldsymbol{\beta}^\top)^\top$.
2: Generate $u \sim \text{Uniform}(0,1)$, then using a numerical method (e.g., `uniroot` function into `stats` package) obtain the lifetime y from the equation

$$\left(1+\frac{\rho(e^{\kappa y}-1)\xi\theta(\theta+4)}{2\kappa(\theta+2)}\right)^{-\frac{4}{\theta(\theta+4)}-1}\left(1+\frac{\rho(e^{\kappa y}-1)\xi\theta}{2\kappa}\right)=u.$$

3: Draw $c \sim \text{Uniform}(0,\varepsilon)$, and compute $t = \min(y,c)$.
4: If $t = y$, then $\nu = 1$, otherwise $\nu = 0$.
5: Repeat the previous steps to obtain the desired sample size.

As shown in Step 3 of the algorithms above, the censoring times were generated from the uniform distribution, where the minimum value is zero and the maximum value ($\varepsilon > 0$) is defined as a way to control the proportion of right-censored observations. The generation scheme also considers covariates. In the following scenarios, we adopted a group indicator covariate: $x = 0$ (control) and $x = 1$ (treatment). We adopted that $X \sim \text{Bernoulli}(p = 0.5)$ for generation of the covariate values.

All simulations were based on 1,000 Monte Carlo runs. We obtained the MLEs by using the `optimr()` function within the `optimx` package (NASH *et al.*, 2020), which is implemented in `R` software (R Core Team, 2021).

4.3.1 Asymptotic properties

In this study, we evaluated the performance of the MLEs of the parameters of the two RWL frailty models considering the following sample sizes: $n = 50$, 100, 300, and 1,000, and censoring proportions: 0%, 10%, 30%, and 50%. For each combination of parameter values, sample size, and censoring proportion, we computed average bias, the standard deviations of the

estimates (SDs), the root mean square errors (RMSEs) of the MLEs of the parameters, and the empirical CPs of 95% CIs.

For the RWL frailty model with Weibull baseline we fixed $\rho = 0.6$, $\kappa = 1.1$, $\theta = 0.7$ and $\beta_1 = 0.7$. While, for the RWL frailty model with Gompertz baseline, we use $\rho = 0.6$, $\kappa = 0.5$, $\theta = 0.8$, and $\beta_1 = 0.7$. The results for the RWL frailty models with Weibull and Gompertz baseline hazard functions are summarized in the Tables 10 and 11, respectively. Overall, the estimation method performed very well for both models. As the sample size increased the bias, RMSE, and SD converged to 0 for all parameters, as expected. Besides, the values of the RMSEs and SDs get close. Empirical CPs for all of the parameters appeared to be reasonably close to the nominal level with increasing sample size, regardless of the censoring rate. For a small sample size ($n \leq 100$), the empirical CPs for the parameters are below the nominal level for some scenarios. These results are also usually observed for other frailty models, for instance, Barker and Henderson (2005) noted the increase of bias and computation issues for samples of size 200 assuming the gamma frailty model. Also, we have observed that as the censoring rate increases, the bias, RMSEs, and SDs increase for a given sample size, as expected. However, when comparing the two studies, we noticed that the metrics RMSE, SD, and CP presented the best performance when we adopted the model with the Weibull baseline hazard function.

4.3.2 Hypothesis testing $H_0 : \theta = 0$

In general, when using frailty models, the interest is in estimating the amount of unobserved heterogeneity present in a sample. In the RWL frailty model with Weibull baseline hazard function, the inclusion of the frailty term is assessed by using the null hypothesis, $H_0 : \theta = 0$. The statistic most commonly used for this purpose is the LR. Asymptotically, it has a chi-squared distribution with one degree of freedom, χ^2_1. However, under H_0, the parameter value is on the boundary of the parametric space, and problems can occur when testing the null hypothesis. Under certain regularity conditions, it has been demonstrated that the statistical distribution of $\Lambda = 2\{\ell(\widehat{\boldsymbol{\Psi}}) - \ell(\widehat{\boldsymbol{\Psi}}_0)\}$, where $\widehat{\boldsymbol{\Psi}}_0$ is the ML estimator under H_0, is a 50/50 mixture of a χ^2_1 distribution and a point mass at 0 (MALLER; ZHOU, 1996).

To assess the behavior of the LR test in testing the null hypothesis, datasets were simulated with different values of sample size and degree of unobserved heterogeneity. For the RWL frailty model with Weibull baseline: we fixed $\rho = 0.6$, $\kappa = 1.1$, $\beta_1 = 0.7$, and $\theta \in \{0, 0.01, 0.10, 0.20, 0.50, 0.75, 1.00, 1.50\}$, thereby simulating arrangements with various amounts of unobserved heterogeneity. Sample sizes were set as: $n \in \{50, 100, 200, 300, 500, 1{,}000\}$. Censored times were generated from the Uniform$(0, 10)$ distribution, with the proportion of censoring times varying from 4% to 19%. For each scenario, the rejection rate of the null hypothesis was calculated. The size and power of the tests are presented in Table 12. As expected, the empirical power of test increases when the sample size increases, and/or when the value of the parameter θ increases. When $\theta \geq 0.50$ and $n \geq 200$, the power of the test is greater

Table 10 – Bias, RMSEs and SDs of the MLEs, and empirical CPs of 95% asymptotic CIs for the simulated data of the RWL frailty model with Weibull baseline hazard function

n		Bias	RMSE	SD	CP	Bias	RMSE	SD	CP
		0% censoring				30% censoring			
50	ρ	−0.0598	0.2660	0.2592	0.9220	−0.0043	0.2670	0.2670	0.8860
	κ	−0.0660	0.2877	0.2801	0.9540	−0.1719	0.4499	0.4157	0.9600
	θ	0.0254	0.3956	0.3948	0.9440	−0.2141	0.7376	0.7058	0.9130
	β_1	−0.0219	0.5676	0.5672	0.9470	−0.1151	0.6831	0.6733	0.9630
100	ρ	−0.0347	0.1703	0.1667	0.9360	−0.0142	0.1848	0.1843	0.9130
	κ	−0.0196	0.1733	0.1722	0.9440	−0.0666	0.2141	0.2035	0.9630
	θ	0.0353	0.2588	0.2563	0.9360	−0.0660	0.4640	0.4592	0.9440
	β_1	−0.0144	0.3628	0.3625	0.9430	−0.0395	0.3689	0.3668	0.9580
300	ρ	−0.0139	0.0934	0.0924	0.9490	−0.0154	0.1128	0.1118	0.9500
	κ	−0.0044	0.0918	0.0917	0.9410	−0.0090	0.1111	0.1107	0.9540
	θ	0.0290	0.1467	0.1438	0.9430	0.0032	0.2642	0.2642	0.9410
	β_1	0.0107	0.2078	0.2075	0.9370	−0.0131	0.2096	0.2092	0.9470
1,000	ρ	−0.0078	0.0485	0.0479	0.9490	−0.0012	0.0579	0.0579	0.9590
	κ	0.0041	0.0490	0.0488	0.9450	−0.0029	0.0582	0.0581	0.9520
	θ	0.0227	0.0788	0.0754	0.9370	0.0001	0.1383	0.1383	0.9470
	β_1	0.0085	0.1055	0.1052	0.9540	0.0050	0.1123	0.1122	0.9470
		10% censoring				50% censoring			
50	ρ	−0.0398	0.2625	0.2594	0.9070	0.0271	0.3250	0.3239	0.8250
	κ	−0.1126	0.3979	0.3816	0.9650	−0.3012	0.6893	0.6199	0.9670
	θ	−0.0694	0.5404	0.5359	0.9580	−0.5813	1.3113	1.1753	0.8300
	β_1	−0.0279	0.6535	0.6529	0.9380	−0.1970	0.8890	0.8669	0.9600
100	ρ	−0.0268	0.1760	0.1739	0.9320	0.0173	0.2029	0.2021	0.8850
	κ	−0.0305	0.1886	0.1862	0.9560	−0.1170	0.3005	0.2767	0.9560
	θ	−0.0014	0.3240	0.3240	0.9310	−0.2513	0.7645	0.7220	0.8840
	β_1	−0.0168	0.3721	0.3717	0.9380	−0.0880	0.4777	0.4695	0.9550
300	ρ	−0.0038	0.0910	0.0909	0.9460	−0.0044	0.1334	0.1334	0.9200
	κ	−0.0188	0.1015	0.0998	0.9530	−0.0306	0.1367	0.1333	0.9470
	θ	−0.0070	0.1703	0.1702	0.9590	−0.0566	0.4297	0.4260	0.9400
	β_1	−0.0054	0.2060	0.2059	0.9400	−0.0304	0.2456	0.2437	0.9400
1,000	ρ	−0.0006	0.0487	0.0487	0.9480	−0.0059	0.0752	0.0749	0.9450
	κ	−0.0014	0.0527	0.0527	0.9390	−0.0052	0.0700	0.0699	0.9540
	θ	0.0038	0.0950	0.0950	0.9470	−0.0036	0.2393	0.2393	0.9420
	β_1	0.0054	0.1106	0.1104	0.9390	−0.0052	0.1244	0.1243	0.9550

Source: Elaborated by the author.

Table 11 – Bias, RMSEs and SDs of the MLEs, and empirical CPs of 95% asymptotic CIs for the simulated data of the RWL frailty model with Gompertz baseline hazard function.

n		Bias	RMSE	SD	CP	Bias	RMSE	SD	CP
		0% censoring				30% censoring			
50	ρ	0.0316	1.2150	1.2146	0.7960	0.1020	0.6801	0.6724	0.7240
	κ	−2.1256	5.9750	5.5837	0.9550	−3.8753	9.6354	8.8209	0.9380
	θ	−1.1258	2.4923	2.2233	0.8880	−1.8748	3.3640	2.7925	0.8350
	β_1	−0.4093	1.1819	1.1087	0.9380	−0.6249	1.5812	1.4523	0.9310
100	ρ	0.0537	0.4879	0.4850	0.8650	0.1057	0.2603	0.2378	0.8310
	κ	−1.1950	3.7216	3.5243	0.9440	−1.8679	4.9085	4.5388	0.9530
	θ	−0.7061	1.7635	1.6158	0.8930	−1.1024	2.1773	1.8773	0.8720
	β_1	−0.2371	0.6666	0.6229	0.9600	−0.3540	0.7536	0.6652	0.9590
300	ρ	0.0249	0.1129	0.1101	0.9270	0.0544	0.1310	0.1192	0.9210
	κ	−0.2726	0.9365	0.8959	0.9240	−0.5662	1.5760	1.4707	0.9660
	θ	−0.1955	0.7389	0.7125	0.9240	−0.3965	1.0488	0.9709	0.9250
	β_1	−0.0717	0.2718	0.2622	0.9500	−0.1400	0.3322	0.3013	0.9660
1,000	ρ	0.0090	0.0564	0.0557	0.9440	0.0091	0.0627	0.0620	0.9590
	κ	−0.0769	0.3266	0.3174	0.9410	−0.0942	0.4159	0.4051	0.9670
	θ	−0.0600	0.3431	0.3378	0.9390	−0.0718	0.4094	0.4031	0.9590
	β_1	−0.0234	0.1440	0.1421	0.9470	−0.0253	0.1498	0.1477	0.9550
		10% censoring				50% censoring			
50	ρ	0.0883	0.6221	0.6158	0.7770	0.1110	0.6504	0.6408	0.7320
	κ	−2.6427	6.8031	6.2683	0.9520	−5.0985	12.0614	10.9296	0.9230
	θ	−1.3486	2.7344	2.3783	0.8710	−2.3711	3.9927	3.2116	0.8030
	β_1	−0.3914	1.1847	1.1182	0.9530	−0.6684	1.6334	1.4902	0.9550
100	ρ	0.0660	0.3080	0.3009	0.8650	0.1470	0.2935	0.2539	0.7830
	κ	−1.4032	4.2897	4.0535	0.9550	−2.9580	6.3017	5.5636	0.9120
	θ	−0.7872	1.9220	1.7532	0.9010	−1.6751	2.8228	2.2715	0.8170
	β_1	−0.2733	0.7595	0.7086	0.9510	−0.4685	0.9488	0.8249	0.9360
300	ρ	0.0264	0.1138	0.1106	0.9260	0.0667	0.1514	0.1359	0.9020
	κ	−0.3116	0.9951	0.9450	0.9420	−0.9323	2.8850	2.7301	0.9510
	θ	−0.2225	0.7873	0.7552	0.9210	−0.6520	1.3693	1.2039	0.9120
	β_1	−0.0701	0.2735	0.2644	0.9470	−0.1759	0.3898	0.3478	0.9590
1,000	ρ	0.0078	0.0574	0.0569	0.9550	0.0168	0.0732	0.0712	0.9540
	κ	−0.0626	0.3229	0.3168	0.9250	−0.1834	0.5459	0.5141	0.9760
	θ	−0.0448	0.3443	0.3414	0.9260	−0.1561	0.5225	0.4986	0.9650
	β_1	−0.0179	0.1425	0.1413	0.9530	−0.0457	0.1603	0.1536	0.9730

Source: Elaborated by the author.

than 0.9. When the null hypothesis is true, the rejection rate is close to the nominal significance level.

Table 12 – Rejection rates of the null hypothesis (absence of unobservable heterogeneity) at 5% nominal significance level for several unobserved heterogeneity and sample sizes considering the RWL frailty model with Weibull baseline hazard function.

	θ							
n	0	0.01	0.10	0.20	0.50	0.75	1.00	1.50
50	0.072	0.093	0.135	0.203	0.457	0.633	0.706	0.799
100	0.059	0.066	0.126	0.281	0.680	0.859	0.929	0.965
200	0.059	0.055	0.187	0.416	0.900	0.988	0.998	1.000
300	0.043	0.062	0.227	0.554	0.969	0.997	1.000	1.000
500	0.048	0.062	0.301	0.721	1.000	1.000	1.000	1.000
1,000	0.039	0.053	0.476	0.927	1.000	1.000	1.000	1.000

Source: Elaborated by the author.

For the RWL frailty model with Gompertz baseline we set $\kappa = 0.5$, $\rho = 0.6$, $\beta_1 = 0.7$ and $\theta \in \{0, 0.01, 0.10, 0.20, 0.50, 0.75, 1.00, 1.50\}$. The sample size was configured to study the model with the Gompertz baseline hazard function. The censored times were generated from the Uniform$(0, 14)$ distribution, with the proportion of censoring times varying from 5% to 17%. The results are shown in Table 13. Again, note that the results are similar to the model with the Weibull baseline hazard function. However, to obtain test power greater than or equal to 0.9, $\theta \geq 0.75$ and $n \geq 500$ are required.

Table 13 – Rejection rates of the null hypothesis (absence of unobservable heterogeneity) at 5% nominal significance level for several unobserved heterogeneity and sample sizes considering the RWL frailty model with Gompertz baseline hazard function.

	θ							
n	0	0.01	0.10	0.20	0.50	0.75	1.00	1.50
50	0.067	0.058	0.087	0.104	0.188	0.334	0.536	0.795
100	0.039	0.037	0.075	0.099	0.254	0.464	0.708	0.961
200	0.038	0.035	0.072	0.136	0.335	0.647	0.899	0.998
300	0.033	0.038	0.076	0.156	0.410	0.741	0.964	1.000
500	0.027	0.032	0.103	0.188	0.527	0.898	0.999	1.000
1,000	0.028	0.026	0.142	0.321	0.769	0.993	1.000	1.000

Source: Elaborated by the author.

4.4 Application on lung cancer data

In this section, we illustrate the applicability of the proposed frailty model by analyzing a real population-based lung cancer dataset. The results obtained from the RWL frailty model with Weibull and gamma baseline distributions were compared to the gamma, BS, and IG frailty

models, and the standard Cox PH model (no frailty). For details about these other frailty models, the readers are referred to the book by Wienke (2010), and the paper by Leão *et al.* (2017). We consider the Weibull and Gompertz baseline hazard functions in all fitted models. For each fitted model, we provide the point estimates and their respective SEs. To select the best model among all fitted models to the data, the AIC, BIC and Bayes factor (BF) are provided. The BF is a useful tool to evaluate the magnitude of the difference between two BIC values. The decision about the best fit is made taking into account the interpretation of twice the natural logarithm of the BF; see (KASS; RAFTERY, 1995; VILCA *et al.*, 2011). The comparisons are performed between models with the same baseline hazard function. Finally, we present an analysis of the Cox-Snell residuals to assess the goodness-of-fit of the selected best model.

The lung cancer dataset is from a retrospective survey of 25,971 records of patients diagnosed with lung in the state of São Paulo, Brazil, between 2000 and 2014. The follow-up of these patients was conducted until 2018. All of the patients were diagnosed with malignant neoplasm of bronchus and lung (C34 - ICD-10 diagnosis code)[1], and both clinical stage III and IV (metastatic) cases were included in the sample. The dataset was provided by the São Paulo Oncocenter Foundation (FOSP), which is responsible for coordinating the Hospital Cancer Registry of the State of São Paulo (<http://fosp.saude.sp.gov.br>). The FOSP is a public institution connected to the State Health Secretariat, which assists in preparing and implementing healthcare policies in Oncology. These policies serve as an instrument for oncology hospitals to prepare their protocols and improve care practices (ANDRADE *et al.*, 2012).

In this study, death due to cancer was defined as the event of interest. The main goal was to evaluate the impact of covariates such as gender, age at diagnosis, clinical stage, surgery, radiotherapy, and chemotherapy on specific survival time, and also to quantify the degree of unobservable heterogeneity in the data. A descriptive analysis of observed covariates is presented in Table 23. In the cohort examined ($n = 25{,}971$), 16,624 (64.01%) patients were male, 10,496 (40.41%) patients were younger (≤ 60 years old), and 16,771 (64.58%) patients were clinical stage IV. Regarding treatment, 3,143 (12.10%) patients underwent surgery, 10,509 (40.46%) of them received radiotherapy, and 7,896 (30.40%) of them received chemotherapy. A total of 24,279 (93.49%) events occurred during the follow-up period. The maximum observation time was approximately 18.75 years, while the median follow-up time was 6.28 years.

Figure 22 presents the KM estimate of the survival function for the lung cancer dataset. According to the estimated curve, the survival rate appears to trend reasonably close to 0 when the time is large. The median lung specific-survival period was approximately 0.620 years. The 2-, 5-, and 10-year specific survival rates are 0.154, 0.044, and 0.021, respectively.

Table 15 shows the summaries of the fitted frailty models with the Weibull and Gompertz baseline hazard functions by considering all the observed covariates. Among the independent

[1] ICD-10 is the 10^{th} revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO).

Table 14 – Descriptive analysis of the observed covariates from the lung cancer dataset.

Covariate	Code	Category	Number of patients ($n = 25{,}971$)	%
X_1: Gender	0	Male	16,624	62.62%
	1	Female	9,347	37.38%
X_2: Radiotherapy	0	No	15,462	59.54%
	1	Yes	10,509	40.46%
X_3: Chemotherapy	0	No	7,896	30.40%
	1	Yes	18,075	69.60%
X_4: Clinical Stage	0	III	9,200	35.42%
	1	IV	16,771	64.58%
X_5: Surgery	0	No	22,828	87.90%
	1	Yes	3,143	12.10%
X_6: Age (years)	0	≤ 60	10,496	40.41%
	1	> 60	15,475	59.59%

Source: Elaborated by the author.

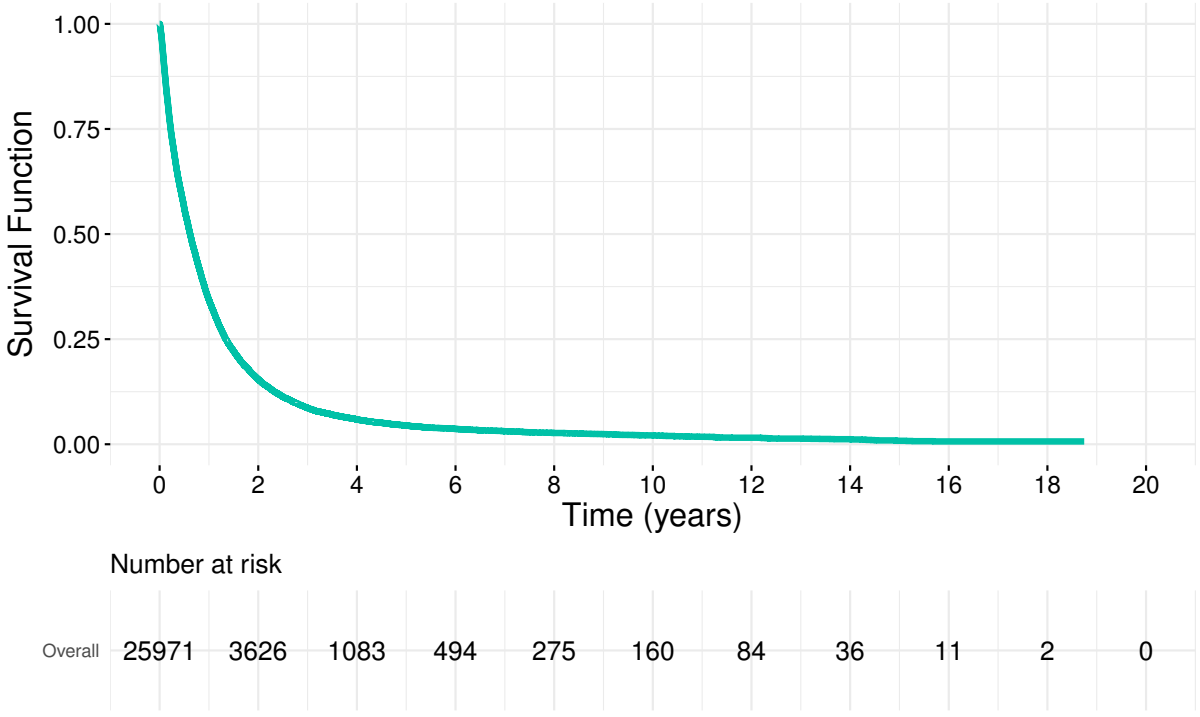

Figure 9 – Estimated survival curve obtained via KM for the lung cancer dataset.

Source: Elaborated by the author.

variables considered in the models, there is evidence that gender, radiotherapy, chemotherapy, clinical stage, and surgery are significant factors in survival time of patients, regardless of the model, since the 95% CIs of the regression coefficients β_j ($j = 1,2,\ldots,5$), calculated by $[\widehat{\beta}_j \pm 1.96 \times \mathrm{SE}(\widehat{\beta}_j)]$, do not include zero. In contrast, age at diagnosis (β_6) was not significant for most models. Despite this, we keep it in the models since it is a clinically significant covariate.

Note that in the fitted frailty models, the estimated amount of unobserved heterogeneity is statistically different from zero (p-values of LR test are less than 0.0001), indicating a degree of unobserved heterogeneity in the data. However, the estimated frailty variances are higher for the frailty models with the Weibull baseline hazard function, inducing more significant heterogeneity among patients.

We report that for the BS and IG frailty models, as well as the Cox PH model, the shape parameter estimates of the Gompertz baseline distribution, $\widehat{\kappa}$, are negative. According to the results, these models indicate that there is a proportion of individuals in the study population that are cured or long-term survivors (MALLER; ZHOU, 1996). The fitted frailty models presented a better fit than the Cox PH models according to the AIC and BIC values, regardless of the baseline hazard function. This result was expected since there is a degree of unobserved heterogeneity in the sample, and the Cox PH model cannot capture it. Besides, the two RWL frailty models have the lowest AIC and BIC values among all fitted models under the same baseline hazard function. Thus, they provide the best fits to the data. This result is also confirmed by means of twice the logarithm of BF values, which indicates a “very strong" evidence in favor of the two RWL frailty models, except when compared to the gamma frailty model with Weibull baseline hazard function. However, this latter BF still provides a “strong" evidence in favor of the RWL frailty model with Weibull hazard function; see (KASS; RAFTERY, 1995; VILCA *et al.*, 2011).

In terms of AIC and BIC, the RWL frailty model with Weibull baseline hazard function provides a better fit than the RWL frailty model with Gompertz baseline hazard function. To assess the goodness-of-fit of these two RWL frailty models, we performed an analysis of the Cox-Snell residuals (COX; SNELL, 1968). The Cox-Snell residuals are defined by:

$$\widehat{e}_i = -\log\left(\widehat{S}(t_i \mid \boldsymbol{x}_i)\right), \quad i = 1,2,\ldots,n, \tag{4.15}$$

where $\widehat{S}(t_i \mid \boldsymbol{x}_i)$ is the estimated survival function of the RWL frailty model with Weibull baseline hazard function for the i^{th} time survival. Thus, when the corresponding RWL frailty model is correctly specified, the Cox-Snell residuals, $\widehat{e}_i$'s, are a censored random sample from the standard exponential distribution (LAWLESS, 2011).

Figure 10 shows exponential quantile-quantile (QQ) plots for Cox-Snell residuals defined in (4.15) based on the two RWL frailty models. We can see that the RWL frailty model with the Weibull baseline hazard function has the best goodness-of-fit for the lung cancer data because the points are close to the identity line. This behavior is not observed using the RWL frailty model with Gompertz baseline hazard function. Therefore, we selected the RWL frailty model with Weibull baseline hazard function as our working model.

Some findings obtained with our working model are given as follows. Initially, observe that the exponential distribution is not supported with $\widehat{\kappa} = 1.601$ and $\text{CI}(\kappa;95\%) = [1.572;1.630]$, which suggests that the baseline hazard function is increasing. The working model shows that the better survival rates are associated with young female clinical stage III

Table 15 – MLEs, SEs, information criteria and twice the logarithm of BF for the fitted frailty and Cox PH models considering the lung cancer dataset. The BF values were calculated assuming the RWL frailty models as correct.

Weibull baseline hazard function					
	RWL	Gamma	BS	IG	Cox PH
Parameter	MLE (SE)	MLE (SE)	MLE (SE)	MLE(SE)	MLE (SE)
κ	1.601 (0.015)	1.642 (0.017)	1.461 (0.015)	1.322 (0.011)	0.902 (0.004)
ρ	0.193 (0.004)	0.186 (0.004)	0.220 (0.006)	0.248 (0.006)	0.482 (0.011)
θ	0.940 (0.020)	1.129 (0.021)	1.650 (0.055)	1.547 (0.082)	–
β_1	−0.295 (0.023)	−0.303 (0.024)	−0.311 (0.022)	−0.290 (0.020)	−0.214 (0.014)
β_2	−0.677 (0.024)	−0.701 (0.025)	−0.574 (0.023)	−0.487 (0.020)	−0.279 (0.013)
β_3	−2.199 (0.034)	−2.250 (0.036)	−1.748 (0.028)	−1.553 (0.024)	−0.928 (0.014)
β_4	0.587 (0.024)	0.605 (0.024)	0.630 (0.023)	0.583 (0.021)	0.431 (0.014)
β_5	−1.024 (0.036)	−1.053 (0.037)	−1.067 (0.034)	−0.989 (0.031)	−0.705 (0.021)
β_6	−0.028 (0.023)	−0.026 (0.023)	0.034 (0.022)	0.034 (0.019)	0.047 (0.013)
$-\ell(\widehat{\boldsymbol{\Psi}})$	20,944.44	20,947.55	21,710.6	21,932.76	23,382.74
AIC	41,906.88	41,913.10	43,439.2	43,883.52	46,781.48
BIC	41,980.36	41,986.58	43,512.68	43,957.00	46,846.80
$2\log(B_{12})$	–	≥ 6	≥ 10	≥ 10	≥ 10

Gompertz baseline hazard function					
	RWL	Gamma	BS	IG	Cox PH
Parameter	MLE (SE)	MLE (SE)	MLE (SE)	MLE (SE)	MLE (SE)
κ	0.036 (0.013)	0.018 (0.011)	−0.089 (0.008)	−0.095 (0.008)	−0.212 (0.005)
ρ	3.470 (0.096)	3.413 (0.093)	3.011 (0.075)	2.985 (0.074)	2.476 (0.051)
θ	0.335 (0.015)	0.320 (0.015)	0.225 (0.015)	0.216 (0.014)	–
β_1	−0.222 (0.017)	−0.222 (0.017)	−0.214 (0.015)	−0.214 (0.015)	−0.199 (0.014)
β_2	−0.374 (0.017)	−0.368 (0.017)	−0.328 (0.015)	−0.325 (0.015)	−0.273 (0.013)
β_3	−1.277 (0.021)	−1.258 (0.020)	−1.119 (0.018)	−1.111 (0.017)	−0.944 (0.014)
β_4	0.439 (0.017)	0.437 (0.017)	0.417 (0.016)	0.415 (0.016)	0.375 (0.014)
β_5	−0.770 (0.026)	−0.765 (0.026)	−0.717 (0.024)	−0.714 (0.024)	−0.637 (0.021)
β_6	0.008 (0.017)	0.010 (0.017)	0.018 (0.015)	0.018 (0.015)	0.020 (0.013)
$-\ell(\widehat{\boldsymbol{\Psi}})$	22,177.02	22,192.32	22,349.21	22,356.36	22,569.46
AIC	44,372.04	44,402.64	44,716.42	44,730.72	45,154.92
BIC	44,445.52	44,476.12	44,789.90	44,804.20	45,220.24
$2\log(B_{12})$	-	≥ 10	≥ 10	≥ 10	≥ 10

Source: Elaborated by the author.

patients who have undergone surgery followed by radiotherapy and chemotherapy. Table 16 shows the estimated 0.5-, 1-, 2-, 3-, and 5-year survival rates according to the RWL frailty model with Weibull baseline hazard function and according to several observed patient characteristics. As expected, female patients exhibited slightly better survival than male patients with the same clinical stage and treatment. Meanwhile, in the absence of treatment, the survival rates are worse, mainly in the later years of diagnosis, as expected.

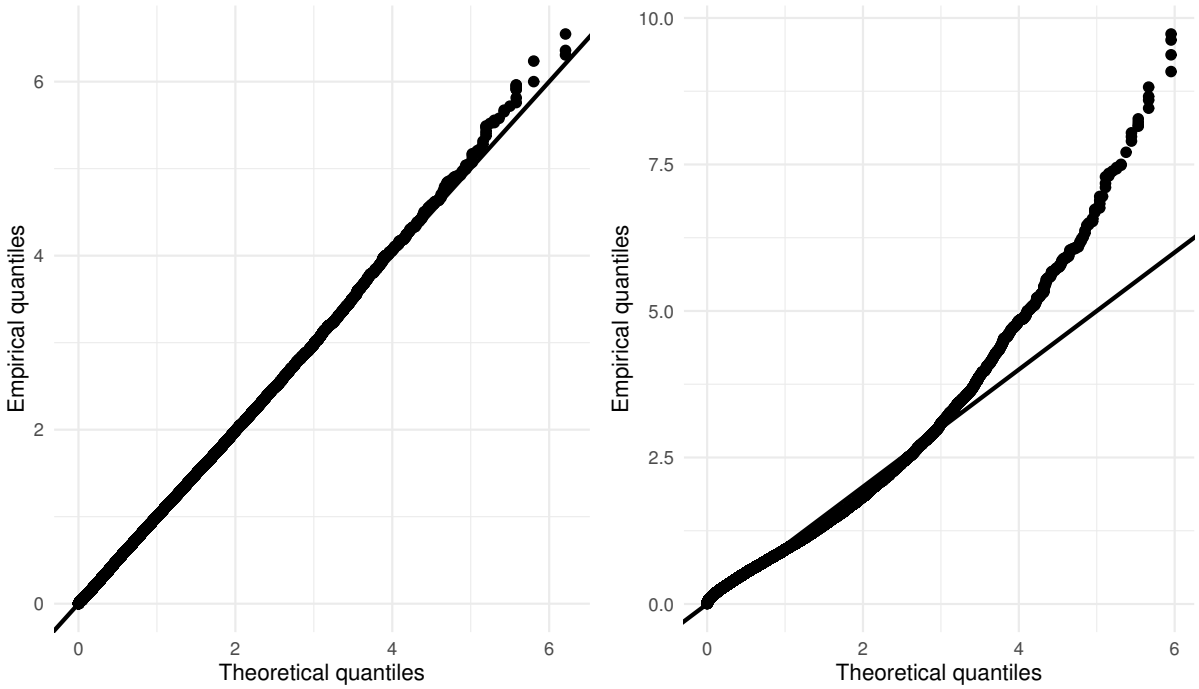

Figure 10 – Exponential QQ plots of Cox-Snell residuals from the RWL frailty models with Weibull (left panel) and Gompertz (right panel) baseline hazard functions for the lung cancer dataset.

Source: Elaborated by the author.

4.5 Concluding remarks

The traditional Cox PH model is not appropriate for survival data in presence of unobserved heterogeneity in a study population. Consequently, its use can lead to erroneous estimates for the regression coefficients. Alternatively, a frailty term (or random effect) must be considered in the Cox PH model to capture unobserved heterogeneity and improve its accuracy. In this chapter, we have proposed a novel frailty model for modeling unobserved heterogeneity in survival data. In this model, the RWL distribution with unitary mean is used as the frailty distribution. When we calculated the Laplace transform of this frailty distribution, both unconditional survival and hazard functions were identified. The Weibull and Gompertz hazard functions were selected as the baseline hazard functions to derive the two RWL frailty models. Monte Carlo simulation studies subsequently showed that the asymptotic properties of the MLEs under different proportions of censoring were satisfied, as expected. Besides, the LR test demonstrated its ability to detect unobserved heterogeneity in both small and large samples, also as expected. However, comparing the simulations for both models, we reported that the RWL frailty model with Weibull baseline hazard function presented better results. We applied our two RWL frailty models to a real lung cancer dataset and compared our results with those given by gamma, BS and IG frailty models, as well as standard Cox PH model. According to AIC, BIC and BF, we reported that the two RWL frailty models returned the best fits to the lung cancer data. Comparing these two selected frailty models, we concluded that the RWL frailty model with Weibull baseline hazard

Table 16 – Estimated specific lung survival rates for older patients stratified by gender, clinical stage and treatment under RWL frailty model.

	No Radiotherapy and No Chemotherapy							
	No Surgery				Surgery			
	CS III		CS IV		CS III		CS IV	
Survival (years)	Male	Female	Male	Female	Male	Female	Male	Female
0.5	0.186	0.232	0.116	0.148	0.381	0.452	0.258	0.316
1.0	0.075	0.096	0.046	0.059	0.174	0.218	0.108	0.138
2.0	0.029	0.038	0.018	0.023	0.070	0.090	0.043	0.055
3.0	0.017	0.022	0.010	0.013	0.041	0.052	0.025	0.032
5.0	0.008	0.011	0.005	0.006	0.020	0.026	0.012	0.016
	Radiotherapy and Chemotherapy							
	No Surgery				Surgery			
	CS III		CS IV		CS III		CS IV	
	Male	Female	Male	Female	Male	Female	Male	Female
0.5	0.798	0.841	0.686	0.746	0.917	0.937	0.859	0.891
1.0	0.564	0.635	0.418	0.491	0.783	0.829	0.667	0.729
2.0	0.301	0.365	0.196	0.245	0.543	0.614	0.398	0.470
3.0	0.187	0.234	0.117	0.149	0.383	0.454	0.259	0.318
5.0	0.097	0.123	0.059	0.076	0.219	0.271	0.138	0.175

Source: Elaborated by the author.

function presented the best fit to the considered dataset. As for some findings, we reported that female patients exhibited slightly better survival than male patients with the same clinical stage and treatment, as expected. Meanwhile, as also expected, the survival rates are worse in the absence of treatment, mainly in the later years of diagnosis.

CHAPTER

5

A LONG-TERM FRAILTY REGRESSION MODEL BASED ON RWL DISTRIBUTION APPLIED TO STOMACH CANCER DATA

Recent advances in medical treatments have increased the interest of researchers in survival models for cancer data incorporating the possibility of immune or cured patients. The observation of an event (e.g., the patient's death) may be due to one or more competing causes. In addition, several unobserved external factors may influence the appearance of a tumor. Our objective in this chapter is to propose a new long-term frailty regression model based on the RWL distribution to jointly accommodate the heterogeneity among patients by their frailties and the presence of a cured fraction of them. The proposed model is found by assuming that the unknown number of competing causes that can influence the survival time follows a negative binomial distribution, which possesses some particular cases. Besides, we suppose that the time for the k-th competing cause to produce the event of interest follows the RWL frailty model with Weibull baseline distribution, given in Chapter 4. A regression structure by considering the logit link function is used to account for the covariate effects in the cure fraction. The proposed cure rate model comprehends the standard mixture, promotion time, and geometric cure rate models. A classical inference is conducted for the parameters of the proposed regression model through ML methods under random right-censoring. Further, we present a Monte Carlo simulation study to verify the MLEs' behavior assuming different sample sizes and censoring proportions. Finally, to illustrate the usefulness of the proposed model, we use it to describe the lifetimes of patients with stomach cancer from a survey conducted in the state of São Paulo, Brazil.

5.1 Model formulation

The proposed long-term frailty regression model is based on the unification of the cure rate models proposed by Rodrigues *et al.* (2009) as follows (see Section 2.5 for details). Let M be

an unobserved non-negative integer-valued random variable denoting the number of competing causes related to the hazard for a sampling unit in the population. In order to take into account for the underdispersion, equidispersion or overdispersion that can be present in the count data, we suppose that M follows a negative binomial (NB) distribution with parameters $\delta \geq -1$ and $\xi > 0$ (denoted by $M \sim \text{NB}(\delta, \xi)$). Then, its probability mass function is given by

$$p_m = P(M = m) = \frac{\Gamma(m + \delta^{-1})}{m!\Gamma(\delta^{-1})} \left(\frac{\delta\xi}{1+\delta\xi}\right)^m (1+\delta\xi)^{-1/\delta}, \quad m = 0, 1, \ldots,$$

where $\Gamma(\cdot)$ is the gamma function and $\delta\xi > -1$. The mean and variance of the NB distribution are $\mathbb{E}[M] = \xi$ and $\mathbb{V}ar[M] = \xi(1+\xi\delta)$, respectively. This parameterization is more useful than the traditional form because the parameter ξ gives the mean number of competing causes, whereas the parameter δ accounts for the inter-individual variance of the number of causes. Besides, the NB distribution includes some well-known distributions as particular cases. In fact, according to Piegorsch (1990), if $\delta = -1/\tau$, for τ positive integer such that $\tau > \xi$, the NB distribution with parameters ξ and $-1/\tau$ gives the same probabilities as a binomial distribution with parameters τ and ξ/τ, that is, $M \sim \text{Bin}(\tau, \xi/\tau)$ for $0 \leq \xi/\tau \leq 1$. As a result, taking $\tau = 1$ implies $\delta = -1$ and we get a Bernoulli distribution with mean ξ. If $\delta \to 0$, we obtain the Poisson distribution with mean ξ, that is, $M \sim \text{Poisson}(\xi)$. Finally, if $\delta = 1$, we have the geometric distribution with parameter $1/(1+\xi)$, that is, $M \sim \text{Geo}(1/(1+\xi))$. Moreover, as pointed out by Castro, Cancho and Rodrigues (2009), if $-1/\xi < \delta < 0$, there is underdispersion from the Poisson model and, on the other hand, if $\delta > 0$, the overdispersion is present. Therefore, we can interpret δ as a dispersion parameter (SAHA; PAUL, 2005).

Given $M = m$, let W_k, $k = 1, 2, \ldots, m$, be IID random variables denoting the time for the k-th competing cause to produce the event of interest. We assume that conditional on M, W_k, $k = 1, 2, \ldots, m$, follows the RWL frailty model with baseline Weibull baseline hazard function without observed covariates. Here, we use the Weibull distribution with shape parameter $\varepsilon > 0$ and scale parameter $\eta > 0$. So that, unconditional survival and hazard functions of W_k's are given, respectively, by

$$S(t \mid \theta, \varepsilon, \eta) = \left[1 + \left(\frac{t}{\eta}\right)^{\varepsilon} \frac{\theta(\theta+4)}{2(\theta+2)}\right]^{-\frac{4}{\theta(\theta+4)}-1} \left[1 + \left(\frac{t}{\eta}\right)^{\varepsilon} \frac{\theta}{2}\right], \quad t > 0, \tag{5.1}$$

and

$$h(t \mid \theta, \varepsilon, \eta) = \varepsilon t^{\varepsilon-1} \left(\frac{4 + \theta(\theta+4)}{2\eta^{\varepsilon}(\theta+2) + t^{\varepsilon}\theta(\theta+4)} - \frac{\theta}{2\eta^{\varepsilon} + t^{\varepsilon}\theta}\right), \quad t > 0, \tag{5.2}$$

where $\theta = 2\left(\phi + \sqrt{\phi(\phi+1)}\right)^{-1}$ is the variance of RWL frailty distribution with unitary mean.

According to Rodrigues *et al.* (2009), the PGF of $M \sim NB(\delta, \xi)$ at $s \in [0, 1]$ is given by

$$G_M(s) = (1 + \delta\xi[1-s])^{-1/\delta}. \tag{5.3}$$

Then, using (5.1) and (5.3), it follows that the population survival function (2.21) becomes

$$S_{pop}(t \mid \theta,\varepsilon,\eta,\delta,\xi) = \left(1+\xi\delta\left\{1-\left[1+\left(\frac{t}{\eta}\right)^{\varepsilon}\frac{\theta(\theta+4)}{2(\theta+2)}\right]^{-\frac{4}{\theta(\theta+4)}-1}\left[1+\left(\frac{t}{\eta}\right)^{\varepsilon}\frac{\theta}{2}\right]\right\}\right)^{-1/\delta}, \tag{5.4}$$

for all $t > 0$. From (5.4), note that $S_{\text{pop}}(t)$ is improper, that is,

$$\lim_{t\to\infty} S_{pop}(t \mid \theta,\varepsilon,\delta,\xi) := p_0 = (1+\delta\xi)^{-1/\delta},$$

with $p_0 > 0$ being the fraction of cured subjects in the population.

The corresponding PDF and hazard function obtained from (5.4) are expressed, respectively, as

$$f_{pop}(t \mid \theta,\varepsilon,\eta,\delta,\xi) = \frac{\xi S(t \mid \theta,\varepsilon,\eta)h(t \mid \theta,\varepsilon,\eta)}{(1+\delta\xi\,[1-S(t \mid \theta,\varepsilon,\eta)])^{\frac{1+\delta}{\delta}}} \tag{5.5}$$

and

$$h_{pop}(t \mid \theta,\varepsilon,\eta,\delta,\xi) = \frac{\xi S(t \mid \theta,\varepsilon,\eta)h(t \mid \theta,\varepsilon,\eta)}{1+\delta\xi\,[1-S(t \mid \theta,\varepsilon,\eta)]}, \tag{5.6}$$

where $S(t \mid \theta,\varepsilon,\eta)$ and $h(t \mid \theta,\varepsilon,\eta)$ are given in (5.1) and (5.2), respectively. Hereafter, we will call the model with survival function (5.4), PDF (5.5) and hazard function (5.6) as Negative Binomial Cure Rate Reparameterized Weighted Lindley Frailty (NBCrRWLF) model. In this model, the frailty parameter θ is used to quantify the unobserved heterogeneity among non-cured subjects.

Figure 12 displays some examples of the shapes obtained for PDF, survival and hazard functions of the NBCrRWLF model when selected values of the parameters were used. With this graphical analysis, it is observed that the PDF and hazard functions present decreasing and unimodal shapes.

5.1.1 Special cases of the NBCrRWLF model

As mentioned before, the NB distribution has, in particular cases, the Bernoulli, Poisson, and geometric distributions. Hence, our NBCrRWLF model reduces to some specific submodels, giving more flexibility to fit real data. These submodels are described as follows.

5.1.1.1 Bernoulli cure rate RWL frailty model

When $\delta = -1$, we get $M \sim \text{Bernoulli}(\xi)$, where $\xi \in [0,1]$ is the probability of success. In this case, the NBCrRWLF model becomes the Bernoulli cure rate RWL frailty (BerCrRWLF)

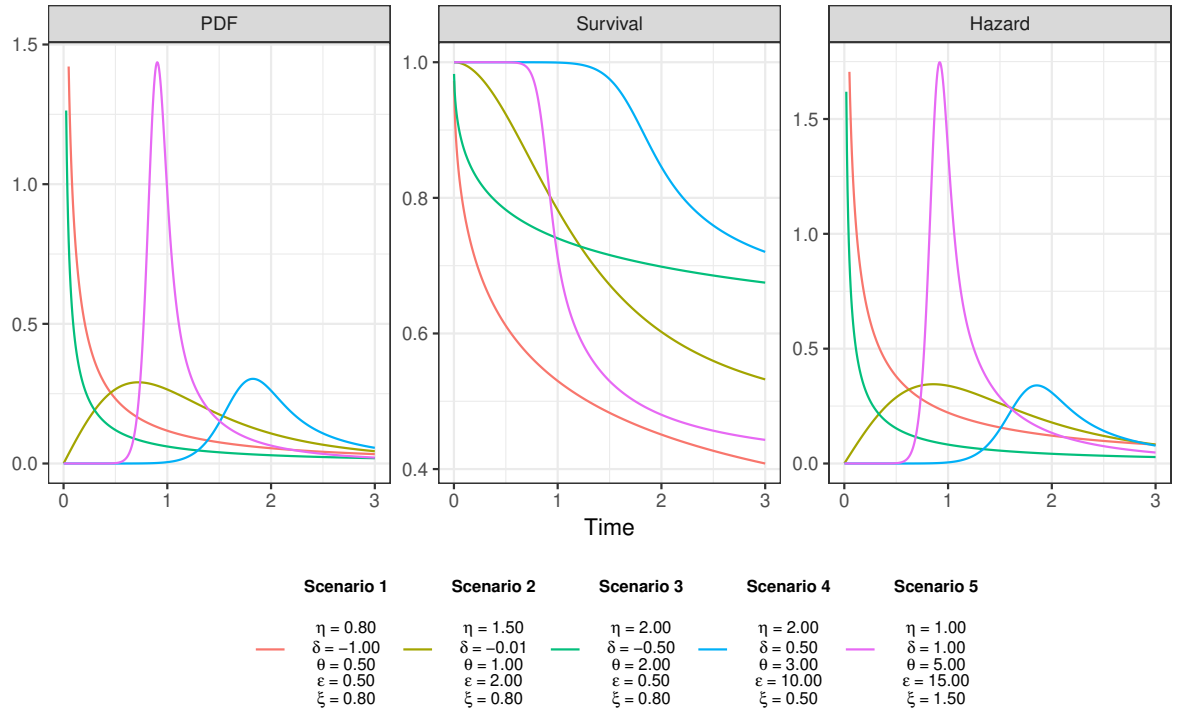

Figure 11 – Some shapes of PDF (right panel), survival (panel middle) and hazard (left panel) functions of the NBCrRWLF model.

Source: Elaborated by the author.

model. The BerCrRWLF long-term survival function is

$$S_{pop}(t \mid \theta,\varepsilon,\eta,\xi) = 1-\xi\left\{1-\left[1+\left(\frac{t}{\eta}\right)^{\varepsilon}\frac{\theta(\theta+4)}{2(\theta+2)}\right]^{-\frac{4}{\theta(\theta+4)}-1}\left[1+\left(\frac{t}{\eta}\right)^{\varepsilon}\frac{\theta}{2}\right]\right\}, \quad (5.7)$$

for all $t > 0$. Notice that there is a relation between the model given in (5.7) and the SMM (BERKSON; GAGE, 1952). In fact, we can rewrite the long-term survival function above as

$$S_{pop}(t \mid \theta,\varepsilon,\eta,\xi) = \pi + (1-\pi)S(t \mid \theta,\varepsilon,\eta),$$

where $\pi = 1-\xi$ and $S(t \mid \theta,\varepsilon,\eta)$ denotes the survival function for the non-cured group in the population as defined in (5.1). In this case, the cure fraction is equal to $p_0 = \pi = 1-\xi$.

5.1.1.2 Poisson cure rate RWL frailty model

If $\delta \to 0$, we obtain $M \sim \text{Poisson}(\xi)$, where $\xi > 0$ is the mean of the number of competing causes. Hence, the NBCrRWLF model reduces to the Poisson cure rate RWL frailty (PoCrRWLF) model, whose long-term survival function is given by

$$S_{pop}(t \mid \theta,\varepsilon,\eta,\xi) = \exp\left\{-\xi\left[1-\left[1+\left(\frac{t}{\eta}\right)^{\varepsilon}\frac{\theta(\theta+4)}{2(\theta+2)}\right]^{-\frac{4}{\theta(\theta+4)}-1}\left[1+\left(\frac{t}{\eta}\right)^{\varepsilon}\frac{\theta}{2}\right]\right]\right\},$$

for all $t > 0$. Notice that we can express the PoCrRWLF model as

$$S_{pop}(t \mid \theta, \varepsilon, \eta, \xi) = \exp\{-\xi\,[1 - S(t \mid \theta, \varepsilon, \eta)]\},$$

where $S(t \mid \theta, \varepsilon, \eta)$ is the survival function for the non-cured group in the population as defined by (5.1). As a result, the PoCrRWLF model can be seen as a particular case of the promotion time cure rate model (YAKOVLEV; TSODIKOV; BASS, 1993), where the cure fraction is given by $p_0 = e^{-\xi}$ for $\xi > 0$.

5.1.1.3 Geometric cure rate RWL frailty model

If $\delta = 1$, we have $M \sim \text{Geo}(1/(1+\xi))$, for $\xi > 0$. Hence, the NBCrRWLF model turns into the geometric cure rate RWL frailty (GeoCrRWLF) model with survival function expressed as

$$S_{pop}(t \mid \theta, \varepsilon, \eta, \xi) = \left(1 + \xi \left\{1 - \left[1 + \left(\frac{t}{\eta}\right)^{\varepsilon} \frac{\theta(\theta+4)}{2(\theta+2)}\right]^{-\frac{4}{\theta(\theta+4)}-1} \left[1 + \left(\frac{t}{\eta}\right)^{\varepsilon} \frac{\theta}{2}\right]\right\}\right)^{-1},$$

for all $t > 0$. Thus, the cure fraction of the GeoCrRWLF model is $p_0 = (1+\xi)^{-1}$ for $\xi > 0$. Figure 12 resumes the particular cases of the NBCrRWLF model through a flowchart.

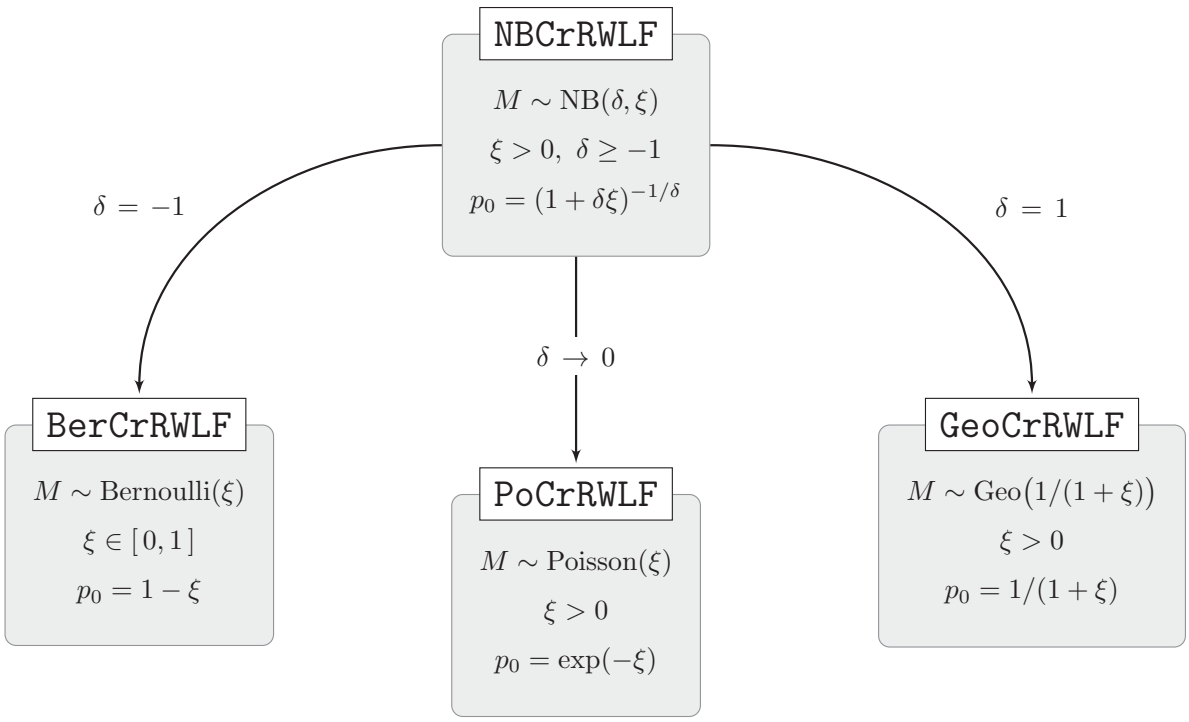

Figure 12 – Some particular cases of the NBCrRWLF model.

Source: Elaborated by the author.

5.1.2 Inference methods

We suppose that some lifetimes are not completely observed for a portion of the individuals and can be subjected to right-censoring. Further, we assume that the censoring is non-informative. Let ν_i denote the censoring indicator variable, taking value 1 if the failure occurs for the i-th individual, and 0 otherwise. Hence, t_i is a lifetime if $\nu_i = 1$, and a censoring time otherwise. Then, for n individuals with observed lifetimes (or censoring times) and their censoring indicators, $(t_1,\nu_1),(t_2,\nu_2),\ldots,(t_n,\nu_n)$ say, which are assumed as independent, the corresponding likelihood function for the parameter vector $\boldsymbol{\Theta} = (\theta,\varepsilon,\eta,\delta,\xi)^\top$ is given by

$$\mathscr{L}(\boldsymbol{\Theta}) = \prod_{i=1}^{n} [f_{pop}(t_i;\boldsymbol{\Theta})]^{\nu_i} [S_{pop}(t_i;\boldsymbol{\Theta})]^{1-\nu_i}, \tag{5.8}$$

where $S_{pop}(\cdot)$ and $f_{pop}(\cdot)$ are the improper survival and PDF functions defined, respectively, in (5.4) and (5.5).

Since the main objective is to estimate the cured fraction in the population, p_0, we put it in the expression of the likelihood function (5.8), using the Fisher's parameterization of the NB distribution (ROSS; PREECE, 1985; CASTRO; CANCHO; RODRIGUES, 2009; LEÃO *et al.*, 2018). For $\delta \geq -1$, we define

$$\xi = -\log(p_0)\,\mathbb{I}(\delta = 0) + \left(\frac{p_0^{-\delta} - 1}{\delta}\right)\mathbb{I}(\delta \neq 0),$$

where $\mathbb{I}(x \in A)$ is the indicator function defined on set A.

Besides, it is well-suited to assume that the cure parameter p_0 could be related to a set of explanatory variables. When these variables are incorporated into the model, we get a different cure rate parameter for each individual, which is denoted by p_{0_i}, for $i = 1,2,\ldots,n$. Now, as $0 \leq p_{0_i} \leq 1$, we can use different link functions to express such a relationship, like probit, logit and log-log, among others; see, e.g., McCullagh and Nelder (MCCULLAGH; NELDER, 1989). In this work, we use the logit link function. Thus,

$$p_{0_i} = \frac{\exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\}}, \quad i = 1,2,\ldots,n,$$

where $\boldsymbol{\beta} = (\beta_0,\beta_1,\ldots,\beta_q)^\top$, for $q < n$, is a vector of regression coefficients to be estimated, which is related to explanatory variables $\boldsymbol{x}_i$, for $i = 1,2,\ldots,n$.

Let $\boldsymbol{\Psi} = (\boldsymbol{\Theta},\boldsymbol{\beta})^\top$ be the model parameters. Then, the log-likelihood function obtained for $\boldsymbol{\Psi}$ is expressed as

$$\ell(\boldsymbol{\Psi}) = \ell_1(\boldsymbol{\Psi})\,\mathbb{I}(\delta = 0) + \ell_2(\boldsymbol{\Psi})\,\mathbb{I}(\delta \neq 0), \tag{5.9}$$

where

$$\ell_1(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \nu_i \log(-\log(p_{0_i})) + \sum_{i=1}^{n} \nu_i \log(h(t_i \mid \theta,\varepsilon,\eta)) + \sum_{i=1}^{n} \nu_i \log(S(t_i \mid \theta,\varepsilon,\eta)) + \sum_{i=1}^{n} \log(p_{0_i}) - \sum_{i=1}^{n} S(t_i \mid \theta,\varepsilon,\eta)\log(p_{0_i})$$

and

$$\ell_2(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \nu_i \log\left(\frac{p_{0_i}^{-\delta}-1}{\delta}\right) + \sum_{i=1}^{n} \nu_i \log\left(h(t_i \mid \theta,\varepsilon,\eta)\right) + \sum_{i=1}^{n} \nu_i \log\left(S(t_i \mid \theta,\varepsilon,\eta)\right) - \sum_{i=1}^{n} \left(\nu_i + \frac{1}{\delta}\right) \log\left(1 + \left(p_{0_i}^{-\delta}-1\right)\left[1 - S(t_i \mid \theta,\varepsilon,\eta)\right]\right),$$

with $S(t \mid \theta,\varepsilon,\eta)$ and $h(t \mid \theta,\varepsilon,\eta)$ being the unconditional survival and hazard functions of T under a RWL frailty as defined in (5.1) and (5.2), respectively.

The MLE $\widehat{\boldsymbol{\Psi}}$ of the parameter vector $\boldsymbol{\Psi}$ can be found by maximizing the log-likelihood function given in (5.9), using some iterative procedure for nonlinear optimization, such as Broyden-Fletcher-Goldfarb-Shanno (BFGS) or quasi-Newton, simulated annealing (SANN), Nelder-Mead, among others (NOCEDAL; WRIGHT, 1999), which are implemented in computational routines of `R` software (R Core Team, 2021). In this work, we adopt the general-purpose unconstrained nonlinear optimization (ucminf) algorithm (NIELSEN; MORTENSEN, 2016).

It is well known in the statistical literature that under certain regularity conditions (LEHMANN; CASELLA, 2006), fulfilled for parameters in the interior of the parameter space but not on the boundary, the MLE $\widehat{\boldsymbol{\Psi}}$ is consistent and follows a normal joint asymptotic distribution with mean $\boldsymbol{\Psi}$ and covariance matrix equal to the $(q+6)\times(q+6)$-inverse of the expected Fisher information matrix $\mathscr{I}(\boldsymbol{\Psi}) = \left\{-\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\Psi})}{\partial\psi_i \partial\psi_j}\right]\right\}$, that is,

$$\widehat{\boldsymbol{\Psi}} \xrightarrow{D} \mathscr{N}_{(q+6)}\left(\boldsymbol{\Psi}, \mathscr{I}^{-1}(\boldsymbol{\Psi})\right) \text{ as } n \to \infty,$$

where $\xrightarrow{D}$ means convergence in distribution. Unfortunately, the exact mathematical expression of the expected Fisher information matrix is difficult to be obtained for the NBCrRWLF model. In this case, we can approximate it by its observed version, defined as $\boldsymbol{H}(\boldsymbol{\Psi}) = \left\{-\frac{\partial^2 \ell(\boldsymbol{\Psi})}{\partial\psi_i \partial\psi_j}\right\}$ evaluated at $\boldsymbol{\Psi} = \widehat{\boldsymbol{\Psi}}$, which can be obtained numerically from the computational routines' results employed. Hence, we can construct approximate $100(1-\rho)\%$ confidence regions for the parameters, as well as hypothesis tests, through the estimated marginal distributions (all normal).

Due to the complexity of our model, the regularity conditions are not easy to check analytically. Therefore, in the next section, we will perform a simulation study to investigate whether the MLEs' usual asymptotic properties hold.

5.2 Simulation study

In this section, we carried out a Monte Carlo simulation study to evaluate the MLEs provided by the NBCrRWLF model under different sample sizes and censoring proportions. For the sake of simplicity, we worked with $\varepsilon = 1$ (baseline exponential) and $\delta \to 0$ (PoCrRWLF

model) fixed. In addition, we divided the sample into two groups (x): control (group 0) and treatment (group 1). Therefore, the cure fraction is computed as

$$p_{0_i} = \begin{cases} p_{00} = \dfrac{\exp(\beta_0)}{1+\exp(\beta_0)}, & \text{if } x_i = 0, \\ \\ p_{01} = \dfrac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}, & \text{if } x_i = 1. \end{cases}$$

We considered a sample size $n \in \{100, 500, 1000, 2500, 5000\}$, with $\theta \in \{0.5, 1, 2\}$, $\eta \in \{1, 1, 1\}$, $p_{00} \in \{0.10, 0.25, 0.90\}$, $p_{01} \in \{0.05, 0.50, 0.65\}$, total censoring proportion $p.cens \in \{0.15, 0.35, 0.75, 0.85\}$, and observed maximum time $t_{max} = 100$. To simulate random samples of size n from the PoCrRWLF model, we used Algorithm 4.

Algorithm 4 – Generator of random numbers from the PoCrRWLF model with $\varepsilon = 1$.

1: Define $\boldsymbol{\Psi} = (\theta, \eta, \beta_0, \beta_1)^\top$;
2: Generate $x_i \sim \text{Bernoulli}(0.5)$;
3: Calculate p_{0_i};
4: Generate $u_i \sim \text{Uniform}(0, 1)$;
5: Use a numerical method to obtain the lifetime y_i from the equation:

$$\exp\left\{\log(p_{0_i})\left[1-\left[1+\frac{\theta(\theta+4)y_i}{2(\theta+2)\eta}\right]^{-\frac{4}{\theta(\theta+4)}-1}\left[1+\frac{\theta y_i}{2\eta}\right]\right]\right\} = u_i;$$

6: If the root y_i in Step 5 was not found, then $y_i = t_{max}$ (censored time due to end of experiment);
7: Draw $y_i^* \sim \text{Uniform}(0, t_{max}^*)$, where t_{max}^* is configured to obtain a total censoring proportion equal to $p.cens$;
8: Determine $t_i = \min(y_i, y_i^*)$. If $t_i = t_{max}$ or $t_i = y_i^*$, then $\nu_i = 0$, otherwise $\nu_i = 1$.
9: Repeat the previous steps to obtain the desired sample size.

All simulations were performed using the R software (R Core Team, 2021), with $N =$ 1,000 Monte Carlo runs. The root in Step 5 was found employing `uniroot` function into `stats` package. Finally, the following performance criteria were considered: MRE, RMSE, and 95% CP, which are computed, respectively, by

$$MRE(\hat{\psi}_i) = \frac{1}{N}\sum_{j=1}^{N}\frac{\hat{\psi}_i^{(j)}}{\psi_i}, \qquad RMSE(\hat{\psi}_i) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\hat{\psi}_i^{(j)} - \psi_i\right)^2},$$

and

$$CP(\hat{\psi}_i) = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}\left(\psi_i \in [a_i^{(j)}, b_i^{(j)}]\right),$$

where ψ_i is the ith component of vector $\boldsymbol{\Psi}$ and $\hat{\psi}_i$ is its associated MLE, $\mathbb{I}(x \in A)$ is the indicator function defined on set A, whereas $a_i^{(j)} = \hat{\psi}_i - 1.96 \times SE(\hat{\psi}_i)$ and $b_i^{(j)} = \hat{\psi}_i + 1.96 \times SE(\hat{\psi}_i)$.

According to these criteria, it is expected that both RMSE and MRE return values closer to zero and one, respectively. Also, we expected that for a large number of experiments using 95% CIs, the relative frequencies of these intervals that covered the true values of parameters should be closer to 0.95 (nominal level).

Figures 13, 14 and 15 shows the empirical MREs, RMSEs, and 95% CPs of the MLEs for each value of the parameters and sample sizes considered under total censoring proportion equal to 0.35, 0.75, and 0.85, respectively. The horizontal dashed lines in this figure correspond to the values of RMSE, MRE, and 95% CP, equal to zero, one, and 0.95, respectively. We observed that the MRE and RMSE of all estimators go to one and zero, respectively, as the sample size increases, meaning that they are asymptotically unbiased and consistent, as expected. On the other hand, as the sample size increases, the 95% CPs of all estimators are close to 0.95 (nominal level), suggesting that all estimators follow an approximately normal distribution, also as expected. Therefore, all these results show that the MLEs' usual asymptotic properties are satisfied, and hence, we conclude that the proposed model returns good results.

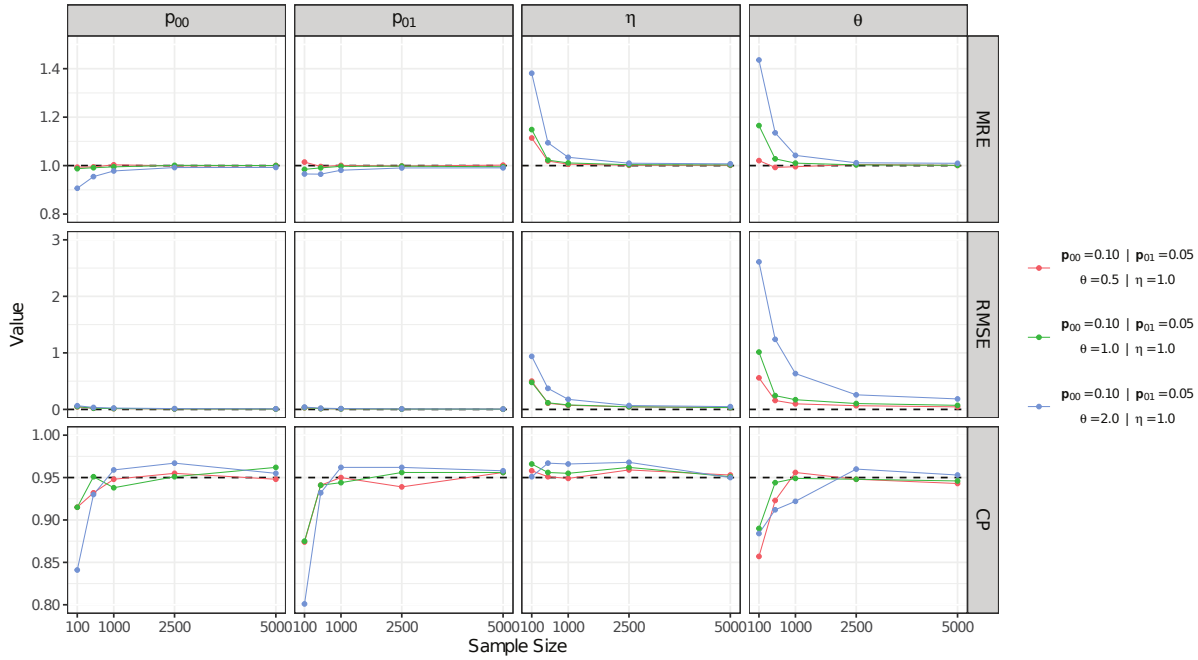

Figure 13 – Empirical MRE, RMSE and 95% CP for the MLEs of θ, η, p_{00} and p_{01} from the PoCrRWLF model, under the indicated n, θ, η, p_{00} and p_{01} values, and also considering $p.cens = 0.35$.

Source: Elaborated by the author.

5.3 Application on stomach cancer patients

This section illustrates the applicability of the proposed model by adopting a new stomach cancer data set. At first, we adjusted the NBCrRWLF model and its special cases. The point and interval estimates are obtained for all models. The estimated survival curves are compared with

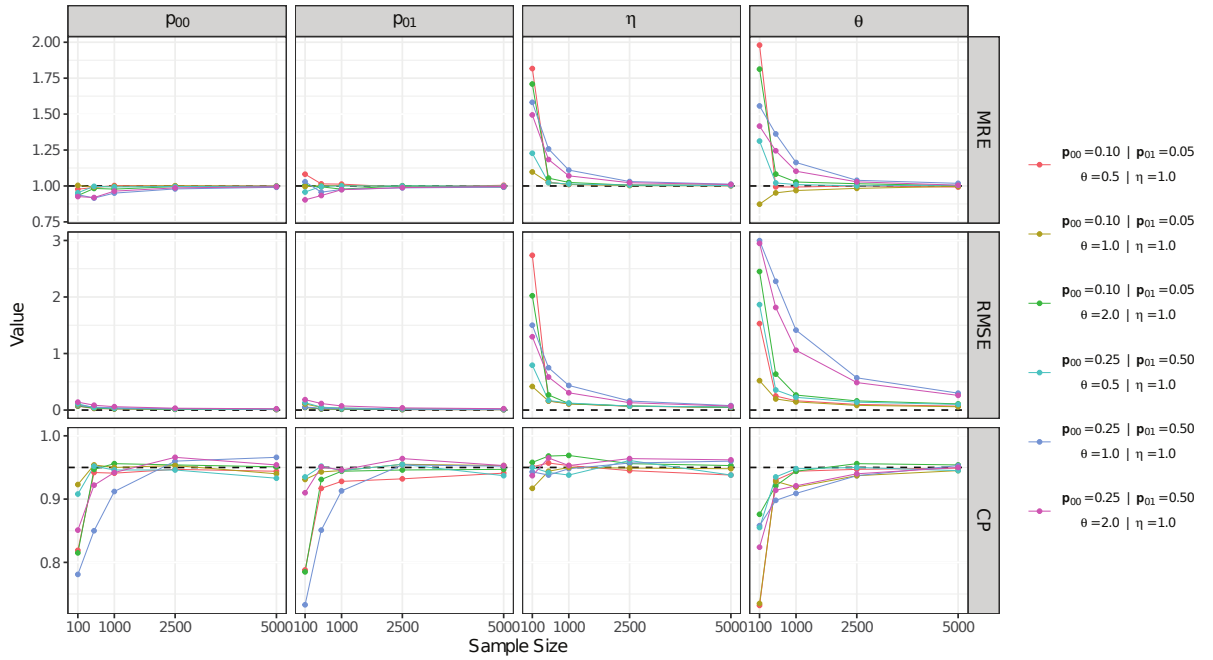

Figure 14 – Empirical MRE, RMSE and 95% CP for the MLEs of θ, η, p_{00} and p_{01} from the PoCrRWLF model, under the indicated n, θ, η, p_{00} and p_{01} values, and also considering $p.cens = 0.75$.

Source: Elaborated by the author.

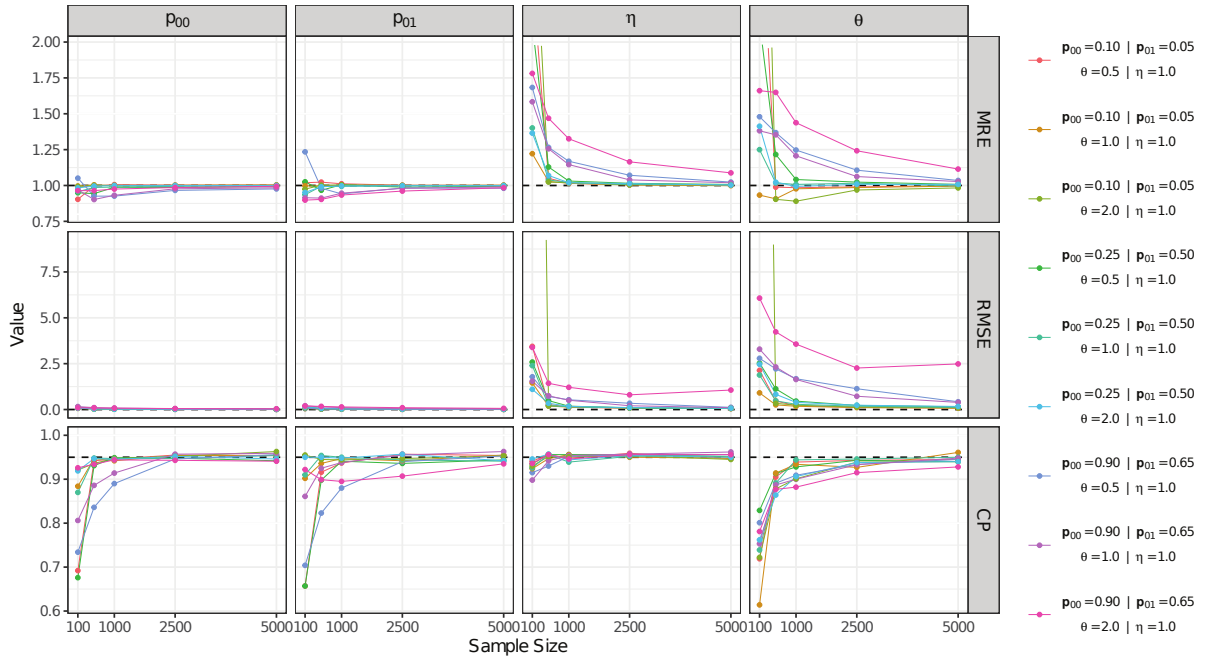

Figure 15 – Empirical MRE, RMSE and 95% CP for the MLEs of θ, η, p_{00} and p_{01} from the PoCrRWLF model, under the indicated n, θ, η, p_{00} and p_{01} values, and also considering $p.cens = 0.85$.

Source: Elaborated by the author.

those obtained by the KM estimator (KAPLAN; MEIER, 1958). The choice of the model that best fits the data is made using the AIC (AKAIKE, 1974). Finally, an analysis of the randomized quantile residuals is also presented for the selected best model by AIC; see (DUNN; SMYTH,

1996).

A retrospective survey of stomach cancer (ICD-10 diagnosis code: C16) patients was obtained from the Fundação Oncocentro de São Paulo (FOSP) (<http://fosp.saude.sp.gov.br/>). The patients included in the study were diagnosed between 2000 and 2014 and were followed-up until 2018. As these are registries from all over the State of São Paulo, 22,148 patients were included in the study. The event of interest is death due to stomach cancer. We observed that 15,065 (68.02%) patients suffered the event of interest, while 7,083 (31.98%) had right-censored times.

Figure 16 shows the survival function obtained by the KM estimator. We note that the 1-, 2-, 5- and 10-years specific survival rates were 0.537, 0.385, 0.277 and 0.246, respectively. Besides, we observe that approximately 21.26% of the patients are long-term survivors.

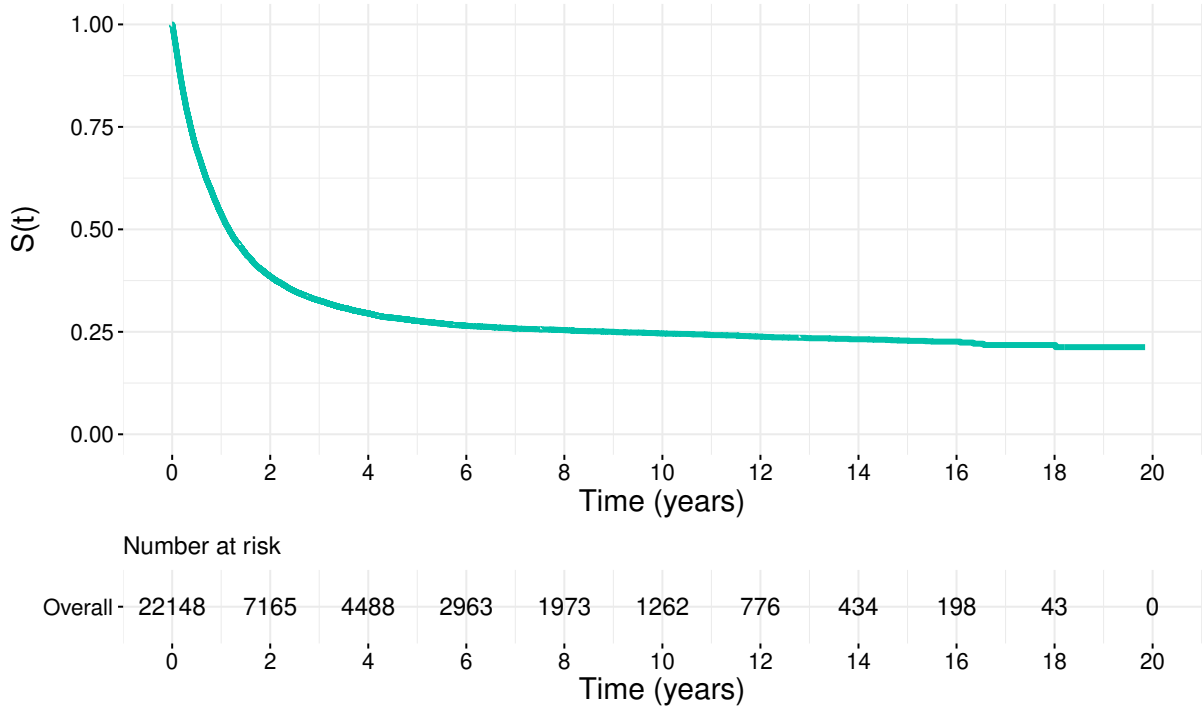

Figure 16 – Estimated survival curve obtained via KM for the stomach cancer data set.

Source: Elaborated by the author.

The main objective is to assess the impact of clinical staging on patient survival and capture the unobserved heterogeneity among the non-cured patients. According to the Brazilian National Cancer Institute (INCA) (<https://www.inca.gov.br/>), staging a cancer case means assessing the tumor's degree of spread. Of the 22,148 patients, 5,744 (25.93%) were classified in clinical stage I or II (group $x = 0$), while 16,404 (74.07%) were classified in clinical stage III or IV (group $x = 1$).

Figure 17 shows the survival functions obtained by the KM estimator for both groups.

We note that patients classified in the group $x = 0$ have a much higher life expectancy than patients in the group $x = 1$. We also note that the percentages of long-term survivors are 54.40% and 9.19% for the groups $x = 0$ and $x = 1$, respectively.

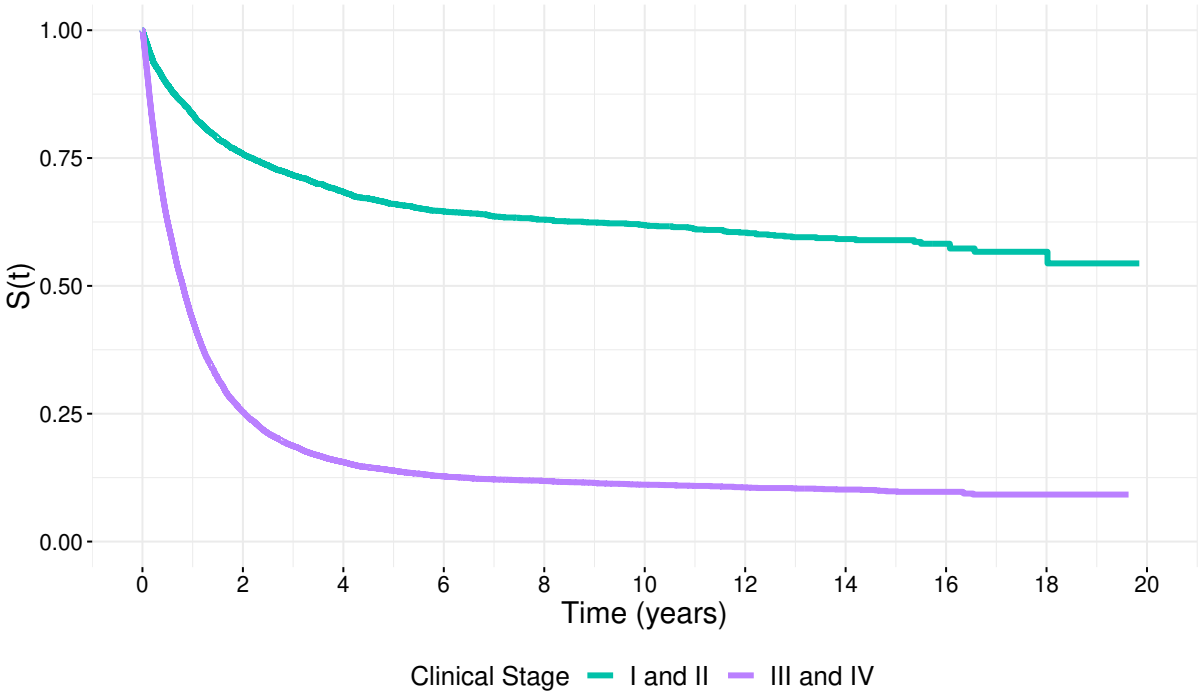

Figure 17 – Estimated survival curve obtained via KM, adopting the clinical stage variable.

Source: Elaborated by the author.

We evaluated the effect of clinical staging by fitting the PoCrRWLF, BerCrRWLF, GeoCrRWLF, and NBCrRWLF models to the data. The results of the fitted models are shown in Table 17. We notice that the NBCrRWLF model's parameter δ is close to zero, indicating that a PoCrRWLF model can be adopted. Analyzing the values of AIC, we reach the same conclusion. We also observe that the PoCrRWLF and NBCrRWLF models' estimates are similar for all parameters in common. Besides, we note that the estimate of the parameter ε is greater than 1, with a 95% CI not containing the value 1, indicating that the exponential distribution cannot be used as baseline distribution.

The results obtained indicate a significant effect of clinical staging since the 95% CI of the parameter β_1 does not include the zero value, for all models. In addition, the estimate of the frailty parameter, θ, is close to 0.65 in the PoCrRWLF, BerCrRWLF, and NBCrRWLF models, while in the GeoCrRWLF model the estimate is 0.339, indicating the existence of unobserved heterogeneity among patients.

In the PoCrRWLF model, the proportions of long-term survivors are: $p_{00} = 0.584$, with 95% CI = $[0.565; 0.603]$, and $p_{01} = 0.081$, with 95% CI = $[0.072; 0.090]$, where p_{00} represents

the proportion of the group $x = 0$, while p_{01} denotes the proportion of the group $x = 1$. Also for the PoCrRWLF model, we tested the hypothesis $H_0 : \theta = 0$ using the LR test. The adopted test statistic considers a correction, because under H_0 the parameter value is on the boundary of the parametric space; for more details, see (MALLER; ZHOU, 1996). We report that the p-value obtained is <0.001, showing that the parameter θ is significant. Hence, other important risk factors influence the patients' lifetimes.

Table 17 – MLE, SE, 95% asymptotic CIs, and AIC value obtained for the PoCrRWLF, BerCrRWLF, GeoCrRWLF and NBCrRWLF models considering clinical stage fitted to the stomach cancer data.

Model	BerCrRWLF				PoCrRWLF			
			95% CI				95% CI	
Parameter	MLE	SE	Lower	Upper	MLE	SE	Lower	Upper
ε	1.094	0.014	1.067	1.121	1.087	0.012	1.063	1.111
η	0.814	0.015	0.785	0.843	1.974	0.045	1.885	2.063
θ	0.646	0.040	0.568	0.724	0.682	0.090	0.505	0.860
β_0	0.509	0.031	0.449	0.569	0.338	0.040	0.260	0.415
β_1 (III and IV)	-3.007	0.066	-3.135	-2.878	-2.767	0.051	-2.867	-2.666
max $\ell(\cdot)$	$-24,995.78$				$-24,804.24$			
AIC	50,001.55				49,618.47			
Model	GeoCrRWLF				NBCrRWLF			
			95% CI				95% CI	
Parameter	MLE	SE	Lower	Upper	MLE	SE	Lower	Upper
δ	–	–	–	–	-0.010	0.050	-0.108	0.088
ε	1.182	0.010	1.162	1.202	1.087	0.012	1.062	1.111
η	3.937	0.129	3.684	4.190	1.956	0.101	1.758	2.153
θ	0.339	0.088	0.167	0.511	0.685	0.091	0.507	0.863
β_0	0.161	0.041	0.081	0.241	0.340	0.041	0.260	0.420
β_1 (III and IV)	-2.207	0.034	-2.273	-2.141	-2.772	0.058	-2.887	-2.658
max $\ell(\cdot)$	$-24,985.91$				$-24,804.22$			
AIC	49,981.81				49,620.43			

Source: Elaborated by the author.

Figure 18 presents the survival curves estimated using the parametric models and the KM estimator. We observe a good approximation between the estimates, except for the BerCrRWLF model for the group $x = 1$. We also note that the proportions of long-term survivors estimated by both methods are close.

Although the fitted survival curves indicate a good fit of the PoCrRWLF model to the data, we will verify such supposition from a residual analysis, which provides us a better view of the fit's quality. The residual analysis is conducted by using the randomized quantile (RQ) residuals, which were proposed by Dunn and Smyth (1996) and are widely used in generalized additive models for location, scale, and shape (GAMLSS); see, e.g., (RIGBY; STASINOPOULOS, 2005). However, in recent years, some applications of these residuals in survival analysis have been developed (YIQI *et al.*, 2016; LEÃO *et al.*, 2018; LOUZADA *et al.*, 2020). Let $\widehat{S}_{\text{pop}}(. \mid \widehat{\Psi})$ be

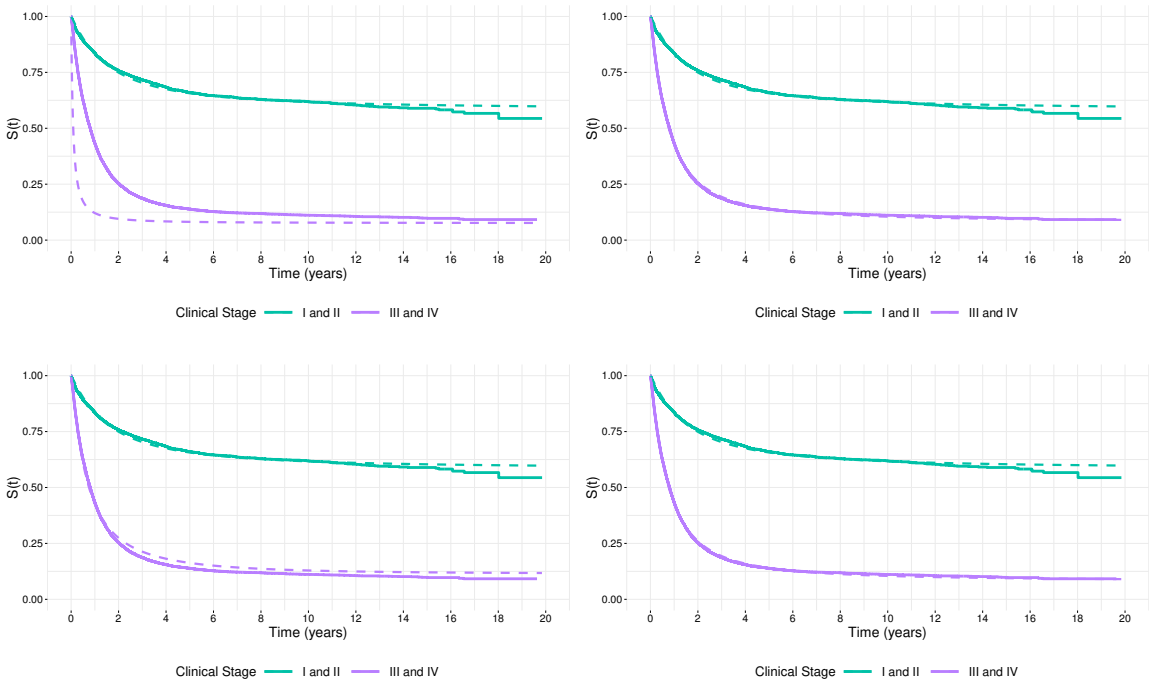

Figure 18 – First row, left to right: Estimated survival curve obtained via KM (solid lines) and BerCrRWLF and PoCrRWLF models (dashed lines), respectively. Second row, left to right: Estimated survival curve obtained via KM (solid lines) and GeoCrRWLF and NBCrRWLF models (dashed lines), respectively.

Source: Elaborated by the author.

the fitted long-term survival function of the PoCrRWLF model. The RQ residuals are defined by

$$\hat{r}_i = \Phi^{-1}\left(\widehat{S}_{\text{pop}}(t_i \mid \widehat{\Psi})\right), \quad i = 1, 2, \ldots, n,$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution. Accordingly, it is expected that the RQ residuals $\hat{r}_i$'s are approximately standard normal if the PoCrRWLF model is correctly specified.

In Figure 19, we present the RQ residuals, along with the 95% confidence limits. Note that the relationship between the theoretical quantiles and the quantile residuals is approximately linear, indicating that the normalized RQ residuals present a good agreement with the standard normal distribution. Therefore, we can consider that the model fitted the data reasonably well.

5.4 Concluding remarks

In this chapter, we have proposed a new long-term frailty regression model based on weighted Lindley distribution. The proposed NBCrRWLF model, came upon by assuming that the unknown number of competing causes that probably affect the survival time follows a negative binomial distribution, absorbing several particular cases. The frailty term was included in the model to quantify the unobserved heterogeneity among non-cured subjects. Besides, we

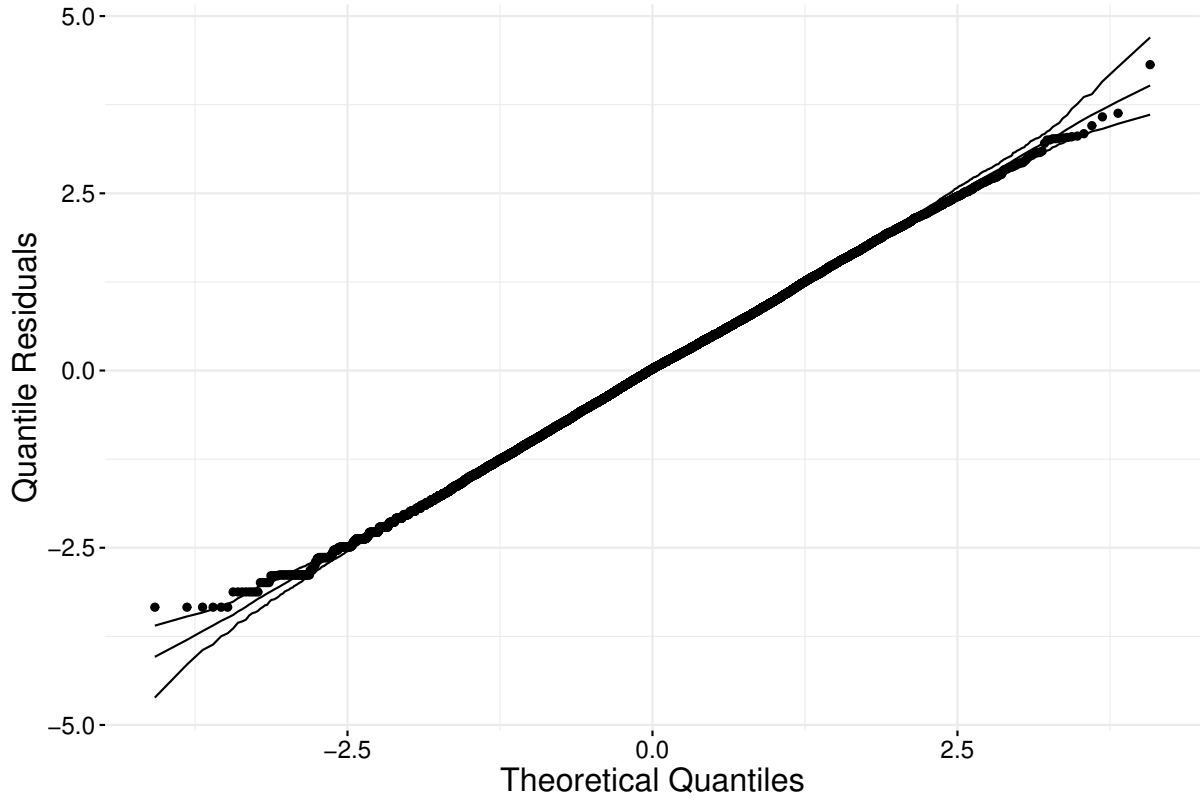

Figure 19 – Plot of normal theoretical quantiles *versus* quantile residuals considering the PoCrRWLF model.

Source: Elaborated by the author.

used the RWL distribution as frailty distribution due to its attractive properties, such as flexibility on its probability density function, Laplace transform on closed-form, among others. We included the cured fraction as a model parameter by using the Fisher's reparameterization of the negative binomial distribution, and it was linked to covariates employing the logit link function. An advantage of the NBCrRWLF model is accommodating some specific compounding cure rate regression models as special cases. Another advantage of the proposed model is the possibility to jointly consider the heterogeneity among patients by their frailties and the presence of a cured fraction of them. A classical inference was conducted for the parameters of the regression model through ML methods under random right-censoring. Monte Carlo simulations showed that the RMSE, MRE, and 95% CP performance criteria returned values reasonably close to their desired levels as sample size increased for all MLEs. Thus, we concluded that the frequentist properties of the MLEs were satisfied, as expected. The proposed model's practical relevance and applicability were demonstrated by describing the lifetime of 22,148 patients with stomach cancer obtained from the Fundação Oncocentro de São Paulo, Brazil. In this real example, we evaluated the clinical staging variable's effect and the frailty variable by fitting the NBCrRWLF model and its special cases (PoCrRWLF, BerCrRWLF, and GeoCrRWLF). The results showed that regression and frailty parameters were statistically significant in all fitted models. Hence, we concluded that the clinical staging variable affects stomach cancer patients' lifetime as well

as other risk factors that were not measured or considered in study planning. We reported that the NBCrRWLF (full) and PoCrRWLF models returned the best fits to the data. We selected the PoCrRWLF model through the selection criterion AIC. Finally, an analysis of RQ residuals indicated a good fit of the PoCrRWLF model to the stomach cancer data.

CHAPTER

6

NON-PROPORTIONAL HAZARDS MODEL WITH A FRAILTY TERM FOR MODELING SUBGROUPS WITH OR WITHOUT EVIDENCE OF LONG-TERM SURVIVORS

NPHs are a common finding in survival analysis. For example, in medical studies, the most common types of NPHs are time-dependent treatment effects, delayed treatment effects, crossing hazards, and diminishing treatment effects over time. In addition, NPHs sometimes also occur due to the random effects. In this chapter, we present a lung cancer dataset with some covariates that exhibit NPHs. In addition, the presence of long-term survivors is observed in subgroups. The proposed framework is based on the GTDL model with the time effect in each subgroup and a random term effect (frailty) to quantify the amount of unobservable heterogeneity. We suppose that the frailty variable follows the RWL distribution with unitary mean. The resulting model allows NPHs and long-term survivors in subgroups. Parameter estimation is performed using the ML method, and Monte Carlo simulation studies are conducted to evaluate the performance of the estimators. We exemplify the use of this model by analysing the survival times of patients diagnosed with lung cancer in the state of São Paulo, Brazil.

6.1 Model formulation

Let $T > 0$ be a random variable representing the failure time and $Z > 0$ be the frailty variable. Using (2.17) and (2.22), we have that the conditional hazard function of the GTDL frailty regression model is defined by

$$h(t \mid z) = z\frac{\lambda \exp(\alpha t + \boldsymbol{x}_1^\top \boldsymbol{\beta})}{1 + \exp(\alpha t + \boldsymbol{x}_1^\top \boldsymbol{\beta})},$$

where $\lambda > 0$ is a scalar, $\alpha \in \mathbb{R}$ is a measure of the time effect, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a vector of p unknown parameters measuring the effect of the p covariates $\boldsymbol{x}_1 = (\boldsymbol{x}_{1_1}, \ldots, \boldsymbol{x}_{1_p})^\top$, and t represents the univariate survival times of the units.

The respective conditional survival function conditional on $Z = z$ is expressed by

$$S(t \mid z) = \left[\frac{1+\exp(\alpha t + \boldsymbol{x}_1^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_1^\top \boldsymbol{\beta})}\right]^{-z\lambda/\alpha}.$$

Milani *et al.* (2015) used the gamma distribution for describing the frailty variable, whereas Calsavara *et al.* (2019a) assumed a PVF frailty distribution. In this work, we suppose that the frailty Z follows the RWL distribution with mean one and shape parameter ϕ, which has PDF given in (4.1). After using the Laplace transform of the PDF frailty (4.2) at $s = \frac{\lambda}{\alpha}\log\left(\frac{1+\exp(\alpha t+\boldsymbol{x}_1^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_1^\top\boldsymbol{\beta})}\right)$, we find that the unconditional survival and hazard functions of the GTDL regression model with RWL frailty (in short, GTDL-RWLF model) are given, respectively, by

$$S(t \mid \lambda, \alpha, \boldsymbol{\beta}, \theta) = \left[1+\log\left(\frac{1+\exp(\alpha t+\boldsymbol{x}_1^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_1^\top\boldsymbol{\beta})}\right)\frac{\lambda\theta(\theta+4)}{2\alpha(\theta+2)}\right]^{-\frac{4}{\theta(\theta+4)}-1} \times \left[1+\log\left(\frac{1+\exp(\alpha t+\boldsymbol{x}_1^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_1^\top\boldsymbol{\beta})}\right)\frac{\lambda\theta}{2\alpha}\right], \quad (6.1)$$

and

$$h(t \mid \lambda, \alpha, \boldsymbol{\beta}, \theta) = \frac{\lambda\exp(\alpha t+\boldsymbol{x}_1^\top\boldsymbol{\beta})\left[4+\theta(\theta+4)\right]}{\left[1+\exp(\alpha t+\boldsymbol{x}_1^\top\boldsymbol{\beta})\right]\left[2(\theta+2)+H_0(t \mid \lambda, \alpha, \boldsymbol{\beta})\theta(\theta+4)\right]} - \frac{\theta\lambda\exp(\alpha t+\boldsymbol{x}_1^\top\boldsymbol{\beta})}{\left[1+\exp(\alpha t+\boldsymbol{x}_1^\top\boldsymbol{\beta})\right]\left[2+H_0(t \mid \lambda, \alpha, \boldsymbol{\beta})\theta\right]}, \quad (6.2)$$

where $H_0(\cdot)$ is the cumulative hazard function of the GTDL model given in (2.18) and $\theta = 2\left(\phi+\sqrt{\phi(\phi+1)}\right)^{-1}$ is the variance of frailty (frailty parameter). Note that the ratio for the hazard functions of two individuals with different covariates $\boldsymbol{x}_{1_1}$ and $\boldsymbol{x}_{1_2}$, $\rho(\boldsymbol{x}_{1_1}, \boldsymbol{x}_{1_2}) = \frac{h_0(t|\lambda,\alpha,\boldsymbol{\beta},\boldsymbol{x}_{1_1},\theta)}{h_0(t|\lambda,\alpha,\boldsymbol{\beta},\boldsymbol{x}_{1_2},\theta)}$, is a function of time, so that the GTDL-RWLF model is of NPHs. Another observation, if the frailty parameter $\theta \to 0$, then the GTDL-RWLF model reduces to traditional GTDL model (2.17).

The unconditional hazard function given in (6.2) can take the decreasing and unimodal forms, see illustration in Figure 20 (right), in addition to the increasing, decreasing and constant forms when $\theta \to 0$. The unimodal form is not observed in the original version of the GTDL model. In Figure 20 (left), we notice several interesting behaviors of the survival function, such as: cure rate in one or both groups and the intersection of the curves. In short, the survival function is proper if $\alpha > 0$ and it is improper for $\alpha < 0$. Therefore, the corresponding long-term survivors proportion is

$$
\begin{aligned}
p(\boldsymbol{x}_1) &= \lim_{t\to\infty} S(t \mid \lambda, \alpha, \boldsymbol{\beta}, \theta) \\
&= \left(1 - \frac{\lambda}{\alpha}\log\left(1+\exp(\boldsymbol{x}_1^\top\boldsymbol{\beta})\right)\frac{\theta(\theta+4)}{2(\theta+2)}\right)^{-\frac{4}{\theta(\theta+4)}-1}\left(1 - \frac{\lambda}{\alpha}\log\left(1+\exp(\boldsymbol{x}_1^\top\boldsymbol{\beta})\right)\frac{\theta}{2}\right) \in (0,1). \qquad (6.3)
\end{aligned}
$$

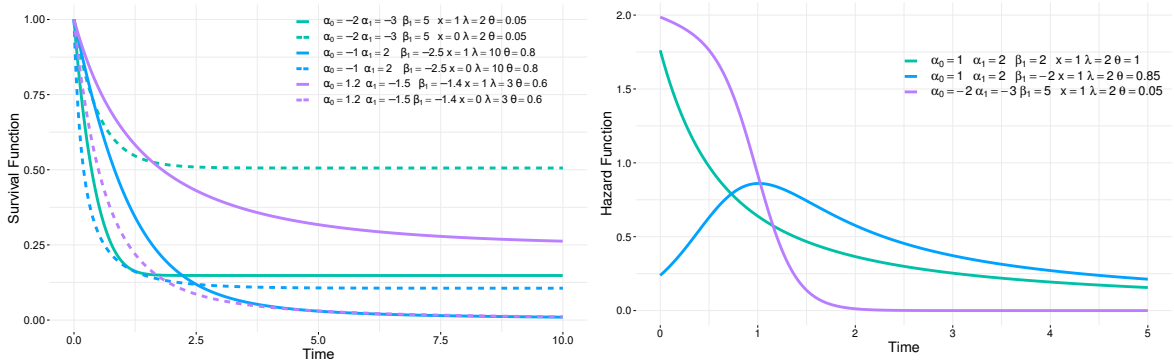

Figure 20 – Unconditional survival (left panel) and hazard (right panel) functions from the GTDL-RWLF model.

Source: Elaborated by the author.

In short, the proposed model has several important characteristics, such as: NPHs; identifies the presence of a long-term survivors without the addition of new parameters; capture the unobserved heterogeneity (if present in the dataset); admit the intersection of survival curves; and allows decreasing and unimodal hazard functions.

6.2 Inference

As done in Calsavara *et al.* (2019a), we also incorporate explanatory variables in the GTDL-RWLF model through parameter α, providing more flexibility to it. For example, when treatment is good, patients can be long-term survivors and the insertion of covariates through this parameter mus lead to estimates less than zero for it. Thus, covariates can be included in the unconditional hazard function (6.2), such as

$$\alpha = \alpha(\boldsymbol{x}_2) = \alpha_0 + \boldsymbol{x}_2^\top\boldsymbol{\alpha}, \qquad (6.4)$$

where α_0 is the intercept and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^\top$ is a vector of q unknown parameters measuring the effect of the covariates vector $\boldsymbol{x}_2 = (x_{2_1}, \ldots, x_{2_q})^\top$. In practice, we can take $\boldsymbol{x}_1 = \boldsymbol{x}_2 = \boldsymbol{x}$, i.e., the covariate vectors may be the same, but we suggest link a subset of variables to the α parameter whether the researcher has prior knowledge about the variables that possibly are associated with the cure rate.

Consider that some lifetimes T_i's are right-censored and we only know that they are greater than the recorded value. Let C_i denote the censoring time, for the ith individual in

study population. Suppose the T_i and C_i are independent random variables. Let $\nu_i = \mathbb{I}(T_i \leq C_i)$ be the censoring indicator, that is, $\nu_i = 1$ if T_i is lifetime and 0 otherwise. We observe $\mathscr{D} = \{(t_i, \nu_i, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}),\ i = 1, \ldots, n\}$, where $t_i = \min(T_i, C_i)$, whereas $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ are the covariates of ith individual. Thus, likelihood function for the parameter vector $\boldsymbol{\Psi} = (\lambda, \alpha_0, \boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \theta)^\top$ assuming non-informative censoring is given by

$$L(\boldsymbol{\Psi} \mid \mathscr{D}) \quad = \quad \prod_{i=1}^{n} \left[h(t_i \mid \boldsymbol{\Psi})\right]^{\nu_i} S(t_i \mid \boldsymbol{\Psi}), \tag{6.5}$$

where $S(\cdot)$ and $h(\cdot)$ are, respectively, the unconditional survival and hazard functions of the GTDL regression model with weighted Lindley frailty defined in (6.1) and (6.2).

Then, taking the natural logarithmic from (6.5), we obtain the log-likelihood function, given by

$$\begin{aligned}\ell(\boldsymbol{\Psi}|\mathscr{D}) = \log(\lambda)\sum_{i=1}^{n}\nu_i + \sum_{i=1}^{n}\nu_i(\alpha(\boldsymbol{x}_{2i})t_i + \boldsymbol{x}_{1i}^\top\boldsymbol{\beta}) - \sum_{i=1}^{n}\nu_i \log\Big(1+\exp(\alpha(\boldsymbol{x}_{2i})t_i + \boldsymbol{x}_{1i}^\top\boldsymbol{\beta})\Big) \\ + \sum_{i=1}^{n}\nu_i \log\left(\frac{4+\theta(\theta+4)}{2(\theta+2)+H_0(t_i)\theta(\theta+4)} - \frac{\theta}{2+H_0(t_i)\theta}\right) \\ - \left(\frac{4}{\theta(\theta+4)}+1\right)\sum_{i=1}^{n}\log\left(1+H_0(t_i)\frac{\theta(\theta+4)}{2(\theta+2)}\right) + \sum_{i=1}^{n}\log\left(1+H_0(t_i)\frac{\theta}{2}\right),\end{aligned} \tag{6.6}$$

where $H_0(\cdot)$ is the GTDL cumulative hazard function given in (2.18) considering the approach described in (6.4), that is

$$H_0(t_i) = \frac{\lambda}{\alpha(\boldsymbol{x}_{2i})}\log\left(\frac{1+\exp(\alpha(\boldsymbol{x}_{2i})t_i + \boldsymbol{x}_{1i}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_{1i}^\top\boldsymbol{\beta})}\right), \quad i = 1, \ldots, n.$$

The estimates of the model parameters can be obtained by maximizing the log-likelihood function (6.6) through numerical methods. In literature, there are several nonlinear optimization algorithms such as best-performing Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton, Nelder-Mead, simulated annealing among other algorithms (NOCEDAL; WRIGHT, 1999). In general, we are also interested in the estimation of long-term survivors. Due to the invariance property of the MLEs (LEHMANN; CASELLA, 2006), the point estimate is obtained using Equation (6.3) and the point estimates of the parameters, while the corresponding SE can be estimated using the delta method.

Let $\widehat{\boldsymbol{\Psi}}$ be the MLE of $\boldsymbol{\Psi}$. Under some standard regularity conditions, the MLE $\widehat{\boldsymbol{\Psi}}$ is consistent and follows a normal joint asymptotic distribution with mean $\boldsymbol{\Psi}$ and covariance matrix $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Psi}})$. Let $\boldsymbol{I}(\boldsymbol{\Psi})$ denote the $(p+q+3)\times(p+q+3)$-expected Fisher information matrix, whose elements are given by

$$\imath_{ij}(\boldsymbol{\Psi}) = -\mathbb{E}_{\boldsymbol{T}|\boldsymbol{\Psi}}\left[\frac{\partial^2\ell(\boldsymbol{\Psi}|\mathscr{D})}{\partial\psi_i\partial\psi_j}\right],$$

where $\mathbb{E}_{\boldsymbol{T}|\boldsymbol{\Psi}}[\cdot]$ means that the expectancy is taken with respect to sample joint distribution of lifetimes. For n large, we have that $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Psi}}) = \boldsymbol{I}^{-1}(\widehat{\boldsymbol{\Psi}})$, where $\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\Psi}})$ is the inverse expected Fisher information matrix (LEHMANN, 2004). However, the expected Fisher information matrix is difficulty to compute with our model. Fortunately, we can approximate it by its observed version, denoted by $\boldsymbol{H}(\widehat{\boldsymbol{\Psi}})$, which is calculated by removing the expectation operator, $\mathbb{E}$, in the expected Fisher information matrix. The advantage of using this observed version is that it can be easily computed by using various computational routines. As a result, we have

$$\widehat{\boldsymbol{\Psi}} \xrightarrow{D} \mathcal{N}_{(p+q+3)}\left(\boldsymbol{\Psi}, \boldsymbol{H}^{-1}(\widehat{\boldsymbol{\Psi}})\right), \quad \text{as } n \to \infty,$$

where $\xrightarrow{D}$ denotes convergence in distribution. Therefore, approximate CIs for each ψ_i with $(1-\gamma)100\%$ are given by

$$\left[\widehat{\psi}_i - z_{\gamma/2}\sqrt{h_{ii}^{-1}(\widehat{\boldsymbol{\Psi}})}; \widehat{\psi}_i + z_{\gamma/2}\sqrt{\widehat{h}_{ii}^{-1}(\widehat{\boldsymbol{\Psi}})}\right],$$

where $\widehat{\psi}_i$ is the MLE of ψ_i, $\widehat{h}_{ii}^{-1}(\widehat{\boldsymbol{\Psi}})$ is the ith diagonal element of inverse observed Fisher information matrix and $z_{\gamma/2}$ denotes the quantile of the standard normal distribution leaving a probability to the right tail with $\gamma/2$. Also, it is possible to conduct hypothesis tests for parameters by using, for example, the ratio likelihood test (LEHMANN; CASELLA, 2006).

Due to the complexity of the GTDL-RWLF model, the regularity conditions are not easy to verify analytically. In this case, simulation studies are required; see, e.g., (HA; MACKENZIE, 2010; ORTEGA *et al.*, 2015; BARRIGA *et al.*, 2019). Following this idea, in the next section, we describe a simulation study performed to investigate whether the usual asymptotic properties of the MLEs hold. We also evaluate the sensitivity of the proposed model in identifying the existence or not of long-term survivors in a subgroup and a sensibility analysis to detect PH when there is proportionality in the data. Besides, we performed a study to evaluate the impact of the MLE caused by the omission of a significant covariate.

6.3 Simulation study

In this section, we report four simulation studies that were carried out to evaluate some properties of the proposed model estimators. In all studies, we evaluated the metrics for different sample sizes and censoring proportions. The simulations were performed with the R Core Team (2021) with 1,000 Monte Carlo runs. We used the delta method with first-order Taylor's approximation to estimate the long-term survivors' SE. The random right-censored data were generated from the exponential distribution with a rate $\eta > 0$, set to control the proportion of right-censored observations. The censoring or failure times of the GTDL-RWLF model were generated using the Algorithm 5, where we divided the sample into two groups (x): control (group 0) and treatment (group 1). We used the same covariate in the two components (α_1 and β_1 parameters).

Algorithm 5 – Generator of random times from the GTDL-RWLF model.

1: Define the values of $\boldsymbol{\Psi} = (\lambda, \alpha_0, \alpha_1, \beta_1, \theta)^\top$.
2: Generate $x \sim \text{Bernoulli}(p)$, with $p \in (0,1)$.
3: $\alpha(x) < 0$, Determine the long-term survivors $p(x)$;
   Generate $u^* \sim \text{Uniform}(0,1)$;
   $u^* < p(x)$, set $t_f = \infty$;
   Generate $u' \sim \text{Uniform}(0, 1-p(x))$;
   Using a numerical method obtain the lifetime t_f from the equation $S(t_f \mid \boldsymbol{\Psi}) = 1-u'$;
   Generate $u \sim \text{Uniform}(0,1)$;
   Using a numerical method obtain the lifetime t_f from the equation $S(t_f \mid \boldsymbol{\Psi}) = 1-u$.
4: Generate $t_c \sim \text{Exponential}(\eta)$.
5: Compute $t = \min(t_f, t_c)$.
6: If $t = t_f$, then $\nu = 1$, otherwise $\nu = 0$.
7: Repeat steps 2 and 6 to obtain the desired sample size.

6.3.1 Asymptotic properties

In this study, we evaluated the performance of the MLEs through a Monte Carlo study. We considered the following sample sizes: $n = 50$, 100, 300, 500, 1,000, 3,000, and 5,000. The metrics adopted in this study were: bias, SDs, MSE, RMSEs, and the CPs of 95% CIs. We denoted p_0 and p_1 as the long-term survivors for the control and treatment groups, respectively. Three scenarios were considered:

(i) without long-term survivors, with $\theta = 0.5$, $\lambda = 1.5$, $\alpha_0 = 0.5$, $\alpha_1 = 0.9$, and $\beta_1 = -1.5$ and censoring proportions of 10%, 20%, and 30%. The results are shown in Table 18;

(ii) without long-term survivors in the control group and with long-term survivors in the treatment group, with $\theta = 0.5$, $\lambda = 1.5$, $\alpha_0 = 0.5$, $\alpha_1 = -0.9$, $\beta_1 = -0.9$, and $p_1 = 0.3732$ and censoring proportions of 30%, 40%, and 50%. The results are shown in Table 19;

(iii) with long-term survivors in both groups, with $\theta = 0.5$, $\lambda = 1.5$, $\alpha_0 = -0.2$, $\alpha_1 = -0.4$, $\beta_1 = -0.9$, $p_0 = 0.0807$, and $p_1 = 0.4921$ and censoring proportions of 40%, 50%, and 60%. The results are shown in Table 20.

We observed that the ML estimates worked very well for the three scenarios because, with an increase in the sample size, the bias, RMSE, MSE, and SD decreased to zero for all parameters. Moreover, the value of the RMSE tended to the SD value with increasing sample size. However, we saw an increase in the bias, RMSE, and SD for fixed sample size when the censoring proportion increased, as expected. Of note, for a sample size less than or equal to 1,000, the empirical CP was sometimes below the nominal level; but for a sample size greater than or equal to 3,000, the empirical CP was close to the nominal level.

Table 18 – Bias, RMSE, SD, MSE, and CP of ML estimates for simulated data considering the GTDL-RWLF model for the scenario (i).

		10% censoring					20% censoring					30% censoring				
n		Bias	RMSE	SD	MSE	CP	Bias	RMSE	SD	MSE	CP	Bias	RMSE	SD	MSE	CP
50	θ	0.0849	0.2468	0.2318	0.2849	0.9650	0.0580	0.2742	0.2680	0.3410	0.9780	0.0355	0.3058	0.3037	0.4245	0.9820
	λ	−0.0362	0.5919	0.5908	0.5984	0.9360	−0.1064	0.6526	0.6438	0.6599	0.9430	−0.1103	0.7139	0.7054	0.7062	0.9420
	α_0	−0.5980	1.7564	1.6514	1.7509	0.8710	−0.7010	1.9438	1.8129	1.9223	0.8970	−0.7995	2.0839	1.9243	2.1696	0.9270
	α_1	−0.0538	2.0382	2.0375	2.4661	0.8710	−0.0859	2.3107	2.3091	2.7681	0.8970	−0.1156	2.5086	2.5060	3.0998	0.9270
	β_1	0.3190	1.0901	1.0423	1.0944	0.9840	0.3676	1.1254	1.0636	1.1435	0.9750	0.3739	1.2245	1.1660	1.2386	0.9740
100	θ	0.0945	0.2132	0.1911	0.1971	0.9090	0.0948	0.2330	0.2128	0.2337	0.9240	0.0779	0.2550	0.2428	0.2978	0.9540
	λ	0.0470	0.4210	0.4183	0.4052	0.8900	0.0479	0.4419	0.4393	0.4229	0.8960	0.0026	0.4812	0.4812	0.4616	0.9110
	α_0	−0.5850	1.8099	1.7127	1.3375	0.8210	−0.5897	1.7580	1.6560	1.4266	0.8640	−0.6652	1.9482	1.8310	1.7237	0.9040
	α_1	0.1342	1.8211	1.8161	1.7616	0.8210	0.0709	1.8722	1.8709	1.8657	0.8640	0.1837	1.9803	1.9717	2.1309	0.9040
	β_1	0.1087	0.7390	0.7310	0.7293	0.9630	0.1285	0.7903	0.7798	0.8033	0.9700	0.1412	0.8518	0.8400	0.8064	0.9660
300	θ	0.0507	0.1276	0.1171	0.1092	0.9250	0.0577	0.1520	0.1406	0.1318	0.9160	0.0727	0.1837	0.1687	0.1635	0.9150
	λ	0.0424	0.2703	0.2669	0.2491	0.9100	0.0374	0.2746	0.2721	0.2570	0.9250	0.0434	0.2784	0.2750	0.2688	0.9180
	α_0	−0.3793	1.4128	1.3608	0.7949	0.8260	−0.4357	1.5599	1.4977	0.9217	0.8300	−0.3590	1.3402	1.2912	0.9302	0.8560
	α_1	0.2473	1.2586	1.2340	0.9046	0.8260	0.3281	1.4127	1.3740	1.0218	0.8300	0.2128	1.3010	1.2835	1.0337	0.8560
	β_1	0.0046	0.4042	0.4042	0.3897	0.9570	−0.0186	0.4196	0.4191	0.4017	0.9420	−0.0119	0.4308	0.4306	0.4254	0.9600
500	θ	0.0310	0.0926	0.0872	0.0833	0.9160	0.0398	0.1141	0.1069	0.0999	0.9240	0.0468	0.1377	0.1295	0.1262	0.9230
	λ	0.0311	0.2111	0.2088	0.1991	0.9290	0.0250	0.2170	0.2155	0.2056	0.9280	0.0282	0.2267	0.2250	0.2155	0.9340
	α_0	−0.2522	1.1165	1.0876	0.5784	0.8700	−0.2546	1.1844	1.1567	0.6453	0.8530	−0.2503	1.0870	1.0578	0.7033	0.8380
	α_1	0.1671	0.9901	0.9758	0.6426	0.8700	0.1964	1.0853	1.0674	0.7010	0.8530	0.2012	0.9865	0.9658	0.7464	0.8380
	β_1	−0.0004	0.2976	0.2976	0.2965	0.9510	−0.0096	0.3147	0.3145	0.3062	0.9510	−0.0180	0.3217	0.3211	0.3199	0.9550
1,000	θ	0.0215	0.0618	0.0579	0.0578	0.9290	0.0202	0.0723	0.0694	0.0689	0.9400	0.0285	0.0950	0.0907	0.0862	0.9250
	λ	0.0211	0.1637	0.1623	0.1444	0.9250	0.0150	0.1613	0.1606	0.1500	0.9370	0.0296	0.1720	0.1695	0.1557	0.9250
	α_0	−0.1316	0.7619	0.7505	0.3691	0.8990	−0.1108	0.6281	0.6182	0.3888	0.8980	−0.1945	0.9755	0.9559	0.4863	0.8900
	α_1	0.1078	0.6905	0.6821	0.4009	0.8990	0.0842	0.5979	0.5919	0.4131	0.8980	0.1523	0.8489	0.8351	0.5001	0.8900
	β_1	−0.0189	0.2097	0.2089	0.2059	0.9460	−0.0113	0.2126	0.2123	0.2121	0.9590	−0.0139	0.2244	0.2240	0.2234	0.9560
3,000	θ	0.0052	0.0325	0.0320	0.0321	0.9490	0.0052	0.0385	0.0381	0.0382	0.9470	0.0080	0.0498	0.0491	0.0485	0.9510
	λ	0.0058	0.0872	0.0870	0.0834	0.9380	0.0046	0.0883	0.0882	0.0858	0.9420	0.0059	0.0920	0.0918	0.0893	0.9510
	α_0	−0.0169	0.1773	0.1765	0.1693	0.9370	−0.0259	0.2852	0.2840	0.1888	0.9430	−0.0279	0.2649	0.2634	0.2176	0.9310
	α_1	0.0076	0.1925	0.1924	0.1881	0.9370	0.0126	0.2836	0.2833	0.2028	0.9430	0.0264	0.2547	0.2533	0.2236	0.9310
	β_1	−0.0049	0.1179	0.1178	0.1165	0.9490	−0.0019	0.1220	0.1220	0.1199	0.9510	−0.0075	0.1242	0.1240	0.1251	0.9540
5,000	θ	0.0019	0.0249	0.0248	0.0247	0.9510	0.0033	0.0292	0.0290	0.0294	0.9480	0.0030	0.0379	0.0378	0.0371	0.9430
	λ	−0.0005	0.0628	0.0628	0.0647	0.9600	0.0029	0.0643	0.0643	0.0661	0.9600	0.0015	0.0703	0.0703	0.0687	0.9460
	α_0	−0.0091	0.1255	0.1251	0.1288	0.9440	−0.0080	0.1398	0.1396	0.1402	0.9460	−0.0147	0.1714	0.1707	0.1611	0.9320
	α_1	0.0030	0.1386	0.1385	0.1430	0.9440	0.0065	0.1516	0.1514	0.1511	0.9460	0.0082	0.1687	0.1685	0.1661	0.9320
	β_1	0.0023	0.0852	0.0851	0.0898	0.9580	−0.0030	0.0899	0.0899	0.0924	0.9560	−0.0016	0.0959	0.0959	0.0962	0.9490

Source: Elaborated by the author.

6.3.2 Sensibility analysis to detect long-term survivors in a subgroup

As previously highlighted, the proposed model can identify long-term survivors' existence, depending on the value of the parameter α. To evaluate the model's sensitivity at identifying the existence or not of long-term survivors, we performed a simulation study considering two groups, control ($x = 0$) and treatment ($x = 1$), with only long-term survivors being observed in the treatment group. Therefore, $p(x = 0) = 0$ and $p(x = 1) \in (0,1)$. To assess the impact of a low probability of long-term survivors, we adopted $p(x = 1) = 0.2, 0.15, 0.1, 0.05$, and 0.03 and different sample sizes. The other parameter values were fixed as $\theta = 0.5$, $\lambda = 1.5$, $\beta_1 = -0.9$, and $\alpha_0 = 0.3$, with the value of α_1 being defined in such a way that the desired long-term survivors were obtained. The proportion of censored data ranged from 33% to 38%. For each simulated dataset, we fitted the model and then checked if $\widehat{\alpha}_0 > 0$ and $\widehat{\alpha}_0 + \widehat{\alpha}_1 < 0$, indicating the absence and presence of long-term survivors in the control and treatment group, respectively. Based on 1,000 Monte Carlo runs, we calculated the percentage of the number of

Table 19 – Bias, RMSE, SD, MSE, and CP of ML estimates for simulated data considering the GTDL-RWLF model for the scenario (ii).

		30% censoring					40% censoring					50% censoring				
n		Bias	RMSE	SD	MSE	CP	Bias	RMSE	SD	MSE	CP	Bias	RMSE	SD	MSE	CP
50	θ	0.0101	0.2892	0.2890	0.4466	0.9940	−0.0594	0.3617	0.3568	0.5661	0.9930	−0.1500	0.5066	0.4839	0.7576	0.9940
	λ	−0.0759	0.6994	0.6952	0.6627	0.9100	−0.1193	0.7554	0.7459	0.7308	0.9300	−0.1889	0.8891	0.8688	0.8430	0.9350
	α_0	−0.7664	1.8501	1.6837	1.4737	0.9060	−0.9343	2.0323	1.8045	2.0357	0.9630	−1.2555	2.4545	2.1087	3.0461	0.9880
	α_1	0.8318	1.9022	1.7106	1.5103	0.9060	0.9889	2.0570	1.8034	2.1067	0.9630	1.2124	2.3451	2.0070	3.1410	0.9880
	β_1	−0.1208	0.8142	0.8052	0.8550	0.9800	−0.1187	0.9817	0.9745	1.0029	0.9750	0.0123	0.9945	0.9945	1.1025	0.9760
	p_0	0.0007	0.1220	0.1220	0.1198	0.9390	0.0303	0.1794	0.1768	0.1663	0.8680	0.0559	0.2278	0.2208	0.2317	0.7440
100	θ	0.0598	0.2311	0.2232	0.3137	0.9700	0.0062	0.2796	0.2795	0.3990	0.9800	−0.0302	0.3381	0.3367	0.5249	0.9910
	λ	0.0259	0.4590	0.4583	0.4380	0.9040	0.0145	0.4695	0.4692	0.4727	0.9100	−0.0051	0.5190	0.5190	0.5215	0.9260
	α_0	−0.5960	1.5995	1.4842	1.1191	0.9120	−0.8080	1.8328	1.6448	1.5499	0.9520	−0.9318	2.0237	1.7962	2.0208	0.9780
	α_1	0.6455	1.6320	1.4987	1.1301	0.9120	0.8465	1.8874	1.6867	1.5651	0.9520	0.9626	2.0859	1.8503	2.0475	0.9780
	β_1	−0.1204	0.6062	0.5941	0.5680	0.9700	−0.1128	0.6526	0.6428	0.6388	0.9690	−0.1121	0.7630	0.7547	0.7300	0.9760
	p_0	−0.0033	0.0816	0.0815	0.0820	0.9500	0.0090	0.1244	0.1241	0.1125	0.9160	0.0396	0.1797	0.1753	0.1642	0.8420
300	θ	0.0605	0.1742	0.1634	0.1923	0.9650	0.0450	0.1961	0.1909	0.2443	0.9620	0.0196	0.2353	0.2345	0.3161	0.9720
	λ	0.0626	0.2786	0.2715	0.2604	0.8910	0.0633	0.2920	0.2850	0.2755	0.9030	0.0708	0.3093	0.3011	0.2985	0.9100
	α_0	−0.4197	1.3890	1.3240	0.7540	0.9000	−0.5100	1.4276	1.3333	0.9454	0.9540	−0.6598	1.5638	1.4176	1.3210	0.9770
	α_1	0.4429	1.4064	1.3347	0.7536	0.9000	0.5329	1.4517	1.3503	0.9423	0.9540	0.6953	1.6082	1.4500	1.3174	0.9770
	β_1	−0.0900	0.3670	0.3558	0.3317	0.9290	−0.1022	0.3994	0.3861	0.3603	0.9370	−0.1315	0.4315	0.4109	0.4078	0.9680
	p_0	0.0007	0.0467	0.0467	0.0471	0.9520	0.0044	0.0606	0.0604	0.0616	0.9600	0.0094	0.0972	0.0967	0.0946	0.9370
500	θ	0.0487	0.1430	0.1345	0.1546	0.9730	0.0541	0.1640	0.1549	0.2019	0.9850	0.0307	0.1914	0.1889	0.2627	0.9800
	λ	0.0394	0.2152	0.2115	0.2123	0.9310	0.0576	0.2339	0.2266	0.2166	0.9250	0.0557	0.2447	0.2383	0.2327	0.9230
	α_0	−0.2686	1.1177	1.0849	0.6233	0.8770	−0.3591	1.1482	1.0905	0.7612	0.9540	−0.4183	1.2072	1.1323	0.9662	0.9870
	α_1	0.2852	1.1260	1.0893	0.6213	0.8770	0.3854	1.1678	1.1023	0.7545	0.9540	0.4415	1.2369	1.1553	0.9528	0.9870
	β_1	−0.0529	0.2650	0.2597	0.2607	0.9470	−0.0912	0.3141	0.3006	0.2810	0.9410	−0.0742	0.3233	0.3147	0.3088	0.9540
	p_0	−0.0003	0.0356	0.0356	0.0363	0.9520	0.0010	0.0474	0.0474	0.0467	0.9460	0.0026	0.0716	0.0716	0.0706	0.9430
1,000	θ	0.0430	0.1075	0.0985	0.1109	0.9740	0.0372	0.1301	0.1247	0.1464	0.9720	0.0263	0.1579	0.1557	0.1992	0.9710
	λ	0.0332	0.1669	0.1636	0.1498	0.9270	0.0338	0.1668	0.1633	0.1566	0.9410	0.0450	0.1841	0.1785	0.1725	0.9250
	α_0	−0.1809	0.9584	0.9412	0.4167	0.8630	−0.1706	0.8418	0.8243	0.5100	0.9170	−0.2970	0.9782	0.9320	0.7379	0.9780
	α_1	0.1907	0.9656	0.9466	0.4137	0.8630	0.1848	0.8502	0.8299	0.5026	0.9170	0.3092	0.9912	0.9417	0.7234	0.9780
	β_1	−0.0332	0.2080	0.2053	0.1831	0.9300	−0.0471	0.2185	0.2134	0.1974	0.9380	−0.0580	0.2403	0.2332	0.2240	0.9400
	p_0	0.0002	0.0263	0.0263	0.0257	0.9460	0.0009	0.0326	0.0326	0.0331	0.9530	0.0039	0.0500	0.0498	0.0488	0.9530
3,000	θ	0.0139	0.0648	0.0633	0.0627	0.9680	0.0187	0.0858	0.0837	0.0888	0.9620	0.0132	0.1156	0.1149	0.1260	0.9530
	λ	0.0096	0.0914	0.0909	0.0908	0.9530	0.0172	0.0994	0.0979	0.0953	0.9510	0.0262	0.1063	0.1030	0.1004	0.9450
	α_0	−0.0280	0.2888	0.2874	0.2503	0.9490	−0.0792	0.5022	0.4959	0.3200	0.9340	−0.1154	0.5017	0.4882	0.4174	0.9660
	α_1	0.0322	0.2881	0.2863	0.2483	0.9490	0.0840	0.5033	0.4962	0.3141	0.9340	0.1213	0.4980	0.4830	0.4042	0.9660
	β_1	−0.0095	0.1092	0.1088	0.1081	0.9530	−0.0148	0.1260	0.1252	0.1174	0.9450	−0.0308	0.1368	0.1333	0.1287	0.9470
	p_0	−0.0006	0.0151	0.0151	0.0148	0.9460	−0.0002	0.0202	0.0202	0.0191	0.9390	0.0013	0.0278	0.0277	0.0277	0.9510
5,000	θ	0.0107	0.0542	0.0531	0.0477	0.9450	0.0139	0.0682	0.0668	0.0688	0.9720	0.0148	0.0957	0.0946	0.1044	0.9560
	λ	0.0037	0.0718	0.0717	0.0701	0.9400	0.0088	0.0700	0.0694	0.0730	0.9590	0.0125	0.0790	0.0780	0.0766	0.9530
	α_0	−0.0149	0.2265	0.2260	0.1881	0.9530	−0.0228	0.2326	0.2315	0.2376	0.9600	−0.0457	0.3529	0.3499	0.3147	0.9500
	α_1	0.0173	0.2233	0.2227	0.1866	0.9530	0.0278	0.2287	0.2270	0.2326	0.9600	0.0500	0.3466	0.3430	0.3022	0.9500
	β_1	−0.0021	0.0840	0.0840	0.0831	0.9400	−0.0093	0.0883	0.0878	0.0901	0.9550	−0.0117	0.1044	0.1038	0.0983	0.9410
	p_0	−0.0003	0.0116	0.0116	0.0115	0.9580	−0.0007	0.0155	0.0155	0.0147	0.9370	0.0003	0.0223	0.0223	0.0215	0.9360

Source: Elaborated by the author.

times the model correctly identified the sign of the parameters. The results are shown in Table 21.

For a fixed proportion of long-term survivors, the percentage of correctly identified signal parameters (the model correctly identifies the long-term survivors in the treatment group and non-immune subjects in the control group) increases with the sample size, as expected. However, for fixed sample size, the percentage decreases as the proportion of long-term survivors decrease mainly for small sample size, indicating a difficulty for the model to correctly identify long-term survivors' presence only in the treatment group. For a sample size of 3,000 or more, the percentage is approximately 90%, regardless of the fixed proportion of long-term survivors.

Table 20 – Bias, RMSE, SD, MSE, and CP of ML estimates for simulated data considering the GTDL-RWLF model for the scenario (iii).

		40% censoring					50% censoring					60% censoring				
n		Bias	RMSE	SD	MSE	CP	Bias	RMSE	SD	MSE	CP	Bias	RMSE	SD	MSE	CP
50	θ	−0.0665	0.4068	0.4013	0.9081	0.9870	−0.0901	0.4804	0.4719	1.2724	0.9780	−0.0854	0.5549	0.5483	1.9487	0.9450
	λ	−0.1739	0.6771	0.6544	0.7430	0.9680	−0.2966	0.8229	0.7675	0.8714	0.9760	−0.3561	0.9117	0.8393	1.0140	0.9890
	α_0	0.0386	0.1583	0.1536	0.4078	1.0000	0.0779	0.2307	0.2171	0.7545	0.9990	0.1699	0.3717	0.3306	1.4232	0.9980
	α_1	0.1150	0.4535	0.4387	0.5195	1.0000	0.2269	0.6894	0.6510	0.8746	0.9990	0.4915	1.0403	0.9167	1.5649	0.9980
	β_1	−0.1523	0.8059	0.7913	0.8884	0.9910	−0.1351	0.9120	0.9019	1.0318	0.9860	−0.2251	0.8977	0.8690	1.2048	0.9970
	p_0	−0.0282	0.0816	0.0765	0.1475	0.9240	−0.0438	0.1071	0.0978	0.2358	0.8700	−0.0747	0.1562	0.1371	0.4208	0.8370
	p_1	0.0013	0.1140	0.1140	0.1161	0.9310	−0.0228	0.1371	0.1352	0.1415	0.9230	−0.0685	0.1602	0.1448	0.1798	0.8700
100	θ	−0.0315	0.3265	0.3250	0.6168	0.9870	−0.0052	0.3693	0.3692	0.8989	0.9790	−0.0333	0.4440	0.4427	1.3481	0.9610
	λ	−0.1224	0.4605	0.4439	0.4958	0.9680	−0.1088	0.4924	0.4802	0.5300	0.9730	−0.1468	0.5069	0.4851	0.6071	0.9740
	α_0	0.0282	0.1339	0.1309	0.2724	0.9980	0.0506	0.1810	0.1737	0.5198	0.9980	0.1004	0.2686	0.2491	0.9511	0.9980
	α_1	0.0261	0.2532	0.2519	0.3093	0.9980	0.0622	0.3506	0.3451	0.5199	0.9980	0.1970	0.6198	0.5877	0.9168	0.9980
	β_1	−0.0249	0.5221	0.5216	0.5549	0.9810	−0.0575	0.5931	0.5903	0.6245	0.9830	−0.1378	0.6858	0.6718	0.7465	0.9900
	p_0	−0.0145	0.0585	0.0567	0.0851	0.9550	−0.0229	0.0800	0.0767	0.1607	0.9140	−0.0487	0.1208	0.1105	0.2913	0.8600
	p_1	0.0019	0.0769	0.0769	0.0811	0.9510	−0.0071	0.0992	0.0989	0.1020	0.9400	−0.0313	0.1188	0.1146	0.1431	0.9110
300	θ	0.0223	0.2455	0.2445	0.3416	0.9900	0.0537	0.2697	0.2643	0.5175	0.9900	0.0517	0.3059	0.3015	0.8223	0.9850
	λ	−0.0099	0.2588	0.2587	0.2589	0.9450	−0.0267	0.2566	0.2552	0.2842	0.9670	−0.0424	0.2708	0.2675	0.3189	0.9800
	α_0	0.0142	0.0991	0.0981	0.1426	0.9970	0.0484	0.1423	0.1338	0.2884	0.9990	0.0764	0.1978	0.1824	0.5761	0.9940
	α_1	0.0041	0.1396	0.1396	0.1593	0.9970	−0.0184	0.1957	0.1948	0.2706	0.9990	0.0098	0.2575	0.2573	0.4940	0.9940
	β_1	−0.0279	0.3235	0.3223	0.3049	0.9480	−0.0102	0.3272	0.3271	0.3306	0.9590	−0.0581	0.3541	0.3493	0.3806	0.9730
	p_0	−0.0027	0.0362	0.0361	0.0450	0.9650	−0.0129	0.0541	0.0525	0.0886	0.9560	−0.0259	0.0785	0.0741	0.1786	0.9390
	p_1	0.0022	0.0460	0.0459	0.0466	0.9490	0.0030	0.0570	0.0569	0.0578	0.9560	0.0007	0.0789	0.0789	0.0870	0.9560
500	θ	0.0359	0.2188	0.2159	0.2620	0.9900	0.0644	0.2502	0.2417	0.3996	0.9950	0.0641	0.2838	0.2765	0.6497	0.9900
	λ	0.0089	0.1894	0.1892	0.1975	0.9500	0.0031	0.1986	0.1986	0.2147	0.9690	−0.0206	0.2089	0.2079	0.2434	0.9710
	α_0	0.0126	0.0888	0.0879	0.1081	0.9910	0.0424	0.1338	0.1269	0.2201	0.9940	0.0832	0.1819	0.1617	0.4464	0.9970
	α_1	0.0033	0.1182	0.1181	0.1207	0.9910	−0.0188	0.1649	0.1638	0.2049	0.9940	−0.0288	0.2142	0.2123	0.3769	0.9970
	β_1	−0.0243	0.2341	0.2328	0.2339	0.9540	−0.0130	0.2502	0.2499	0.2532	0.9520	−0.0248	0.2794	0.2783	0.2887	0.9630
	p_0	0.0002	0.0299	0.0299	0.0342	0.9660	−0.0089	0.0460	0.0451	0.0698	0.9430	−0.0285	0.0694	0.0633	0.1412	0.9570
	p_1	0.0007	0.0367	0.0367	0.0361	0.9530	0.0010	0.0444	0.0444	0.0443	0.9480	−0.0003	0.0618	0.0618	0.0662	0.9640
1,000	θ	0.0074	0.1789	0.1788	0.1798	0.9610	0.0592	0.2219	0.2139	0.2774	0.9980	0.0988	0.2530	0.2330	0.4608	0.9910
	λ	−0.0056	0.1424	0.1423	0.1408	0.9480	0.0155	0.1447	0.1438	0.1496	0.9410	−0.0042	0.1477	0.1476	0.1685	0.9730
	α_0	0.0022	0.0691	0.0691	0.0730	0.9760	0.0293	0.1137	0.1099	0.1518	0.9860	0.0802	0.1663	0.1456	0.3165	0.9930
	α_1	0.0022	0.0806	0.0806	0.0823	0.9760	−0.0116	0.1204	0.1198	0.1416	0.9860	−0.0438	0.1712	0.1655	0.2659	0.9930
	β_1	−0.0078	0.1683	0.1681	0.1647	0.9460	−0.0136	0.1818	0.1813	0.1780	0.9470	0.0008	0.1891	0.1891	0.2000	0.9640
	p_0	0.0013	0.0223	0.0223	0.0230	0.9680	−0.0039	0.0362	0.0359	0.0485	0.9460	−0.0212	0.0565	0.0523	0.0998	0.9550
	p_1	0.0014	0.0254	0.0254	0.0255	0.9490	−0.0003	0.0317	0.0317	0.0312	0.9530	−0.0016	0.0437	0.0437	0.0460	0.9480
3,000	θ	0.0049	0.1056	0.1055	0.1023	0.9400	0.0105	0.1463	0.1459	0.1573	0.9670	0.0701	0.2050	0.1926	0.2710	0.9960
	λ	0.0017	0.0809	0.0809	0.0807	0.9530	−0.0029	0.0833	0.0833	0.0875	0.9660	0.0070	0.0917	0.0914	0.0966	0.9550
	α_0	0.0007	0.0414	0.0414	0.0407	0.9500	0.0028	0.0801	0.0800	0.0875	0.9700	0.0482	0.1356	0.1268	0.1908	0.9720
	α_1	0.0025	0.0468	0.0468	0.0466	0.9500	0.0016	0.0756	0.0756	0.0814	0.9700	−0.0246	0.1252	0.1228	0.1588	0.9720
	β_1	−0.0049	0.1011	0.1010	0.0948	0.9290	−0.0023	0.1000	0.1000	0.1027	0.9560	−0.0059	0.1183	0.1182	0.1152	0.9530
	p_0	0.0007	0.0127	0.0127	0.0126	0.9560	0.0022	0.0261	0.0260	0.0289	0.9620	−0.0105	0.0431	0.0418	0.0641	0.9390
	p_1	−0.0002	0.0142	0.0142	0.0147	0.9570	0.0004	0.0172	0.0172	0.0180	0.9580	−0.0020	0.0264	0.0263	0.0262	0.9470
5,000	θ	0.0034	0.0785	0.0784	0.0790	0.9590	0.0032	0.1229	0.1228	0.1204	0.9510	0.0447	0.1773	0.1716	0.2136	0.9790
	λ	−0.0037	0.0636	0.0635	0.0627	0.9480	0.0012	0.0664	0.0664	0.0676	0.9540	0.0041	0.0728	0.0726	0.0750	0.9570
	α_0	0.0017	0.0311	0.0310	0.0314	0.9600	−0.0009	0.0684	0.0684	0.0662	0.9610	0.0266	0.1175	0.1144	0.1542	0.9620
	α_1	−0.0007	0.0364	0.0364	0.0359	0.9600	0.0029	0.0625	0.0624	0.0619	0.9610	−0.0126	0.1008	0.1000	0.1279	0.9620
	β_1	0.0006	0.0745	0.0745	0.0733	0.9400	−0.0039	0.0784	0.0783	0.0796	0.9530	−0.0056	0.0911	0.0909	0.0892	0.9440
	p_0	0.0001	0.0094	0.0094	0.0097	0.9570	0.0022	0.0221	0.0220	0.0217	0.9540	−0.0040	0.0382	0.0380	0.0528	0.9310
	p_1	0.0004	0.0112	0.0112	0.0114	0.9570	0.0002	0.0137	0.0137	0.0139	0.9510	−0.0002	0.0202	0.0202	0.0204	0.9480

Source: Elaborated by the author.

6.3.3 Sensibility analysis to detect PH

The main assumption of the proposed model is the non-proportionality of the hazards. We carried out a simulation study to analyze the proposed model's behavior when such an assumption is false. For this, we generated failure times from exponential distribution with $\lambda = \exp(\beta_0 + \beta_1 x)$, where x is a group covariate, as previously defined, being generated from the Bernoulli distribution ($p = 0.5$). We fixed $\beta_0 = 0$, $\beta_1 = 0.5$ and considered different sample

Table 21 – Percentage of the number of cases correctly identified by the proposed model when there was long-term survivors in a subgroup.

Long-term survivors fixed	Sample size							
	50	100	300	500	1,000	2,000	3,000	5,000
0.20	76.7	82.5	84.5	84.9	84.1	86.8	87.8	91.7
0.15	73.7	83.7	85.0	85.3	85.2	86.3	88.5	90.3
0.10	71.0	82.4	85.6	85.9	87.4	86.5	90.1	91.0
0.05	68.4	80.8	84.9	87.8	87.8	89.6	90.1	93.0
0.03	69.1	75.3	84.9	85.1	87.9	87.9	90.8	92.1

Source: Elaborated by the author.

sizes. The exponential model is a classic example that satisfies the assumption of PH. Based on 1,000 Monte Carlo runs, we calculated the proportion of times that the CIs of the parameters α_0, α_1, and θ simultaneously contained the value zero, indicating that the hazards are proportional. The results for different censoring rates are shown in Table 22.

Table 22 – Percentage of the number of cases identified with PH using the proposed GTDL-RWLF.

n	0% censoring	15% censoring	30% censoring
50	98.2	97.8	97.1
100	98.9	98.4	98.1
300	98.5	98.3	97.0
500	97.6	98.1	97.0
1000	96.1	96.6	95.8
3000	90.4	92.2	92.9
5000	87.1	88.5	91.0

Source: Elaborated by the author.

For fixed sample size, the percentage of the number of cases indicating PH presents a slight variation when the censoring rates increase. However, when the censoring rate is fixed, the percentage decreased as the sample size increased, as expected, and the point estimate of the parameters α_0, α_1, and θ tend to be close to zero when the sample size increased. Besides, when the sample size is small, higher SEs are expected compared with the large sample size, which can influence the amplitude of the confidence interval and, consequently, in coverage or not of the value 0. According to the simulation study, approximately 90% or more of cases indicated PH, which suggests an excellent performance of the proposed model in identifying PH when the data is proportionality hazards.

6.3.4 Analysis of bias caused by the absence of a covariate

In this simulation study, we are interested in analyzing the behavior of parameter estimates in the absence of an important covariate in the explanation of failure times. For this, we adopted the time generation process with two covariates, X_1 and X_2, both assuming only 0 or 1 values using the Algorithm 1 with an adaptation for two covariates. We can assume, for example, that the variable X_1 refers to gender (female and male), while the variable X_2 is an indication of the type of treatment received (control and treatment). The values of the adopted parameters were: $\theta = 0.5$, $\lambda = 1.5$, $\alpha_0 = -0.2$, $\alpha_1 = -0.25$, $\alpha_2 = -0.55, -0.05, 0.05$ and 0.55, $\beta_1 = -0.5$ and $\beta_2 = -0.9, -0.3, 0.3$ and 0.9.

We considered four different values for α_2 and β_2 in order to simulate covariates with different levels of importance to explain the time to failure. For each combination, datasets with several sample size $n = 50$, 100, 300, 500, 1000, 3000 and 5000 were generated. For each dataset generated, we obtained the MLE of the fitted complete and incomplete model parameters. For the complete model, both covariates were considered, while in the incomplete one, only the covariate X_1 was considered. Based on 1,000 Monte Carlos runs, we calculated the bias of the parameters for both fitted models.

In all the studied scenarios, the bias of the complete model's parameters decreased with the increase of the sample size, as expected. These results have been omitted here. Figure 21 shows the parameter biases of the incomplete model.

According to the results, when α_2 assumes the values 0.05 or -0.05 and $\beta_2 = -0.3$ or 0.3, the estimated biases of the parameters θ, α_0, α_1 and β_1 tend to zero as sample size increase. Regarding parameters λ and β_1, the results suggested that the estimated bias is positive when β_2 is negative, while the estimated bias is negative if β_2 is positive and slight variations were observed when we vary the value of α_2. Therefore, according to the simulation study, there is evidence that the non-inclusion of a covariate with low influence ($|\alpha_2| \leq 0.05$ and $|\beta_2| \leq 0.3$) in the lifetime reflects in slightly biased estimates when the sample size is large.

6.4 Application on lung cancer data

In this section, the applicability of the proposed model is illustrated in a real lung cancer dataset in Brazilian patients. We fitted the GTDL-RWLF model, in addition to the traditional GTDL model to the dataset and compared with survival curve estimates obtained by using the KM estimator kaplan. The MLEs, SEs, 95% CIs estimates for the parameters, and AIC values (AKAIKE, 1974) were determined for each fitted model.

The data are part of a study about lung cancer comprising 30,900 records of patients diagnosed with lung cancer in the State of São Paulo, Brazil, between 2,000 and 2,014, with

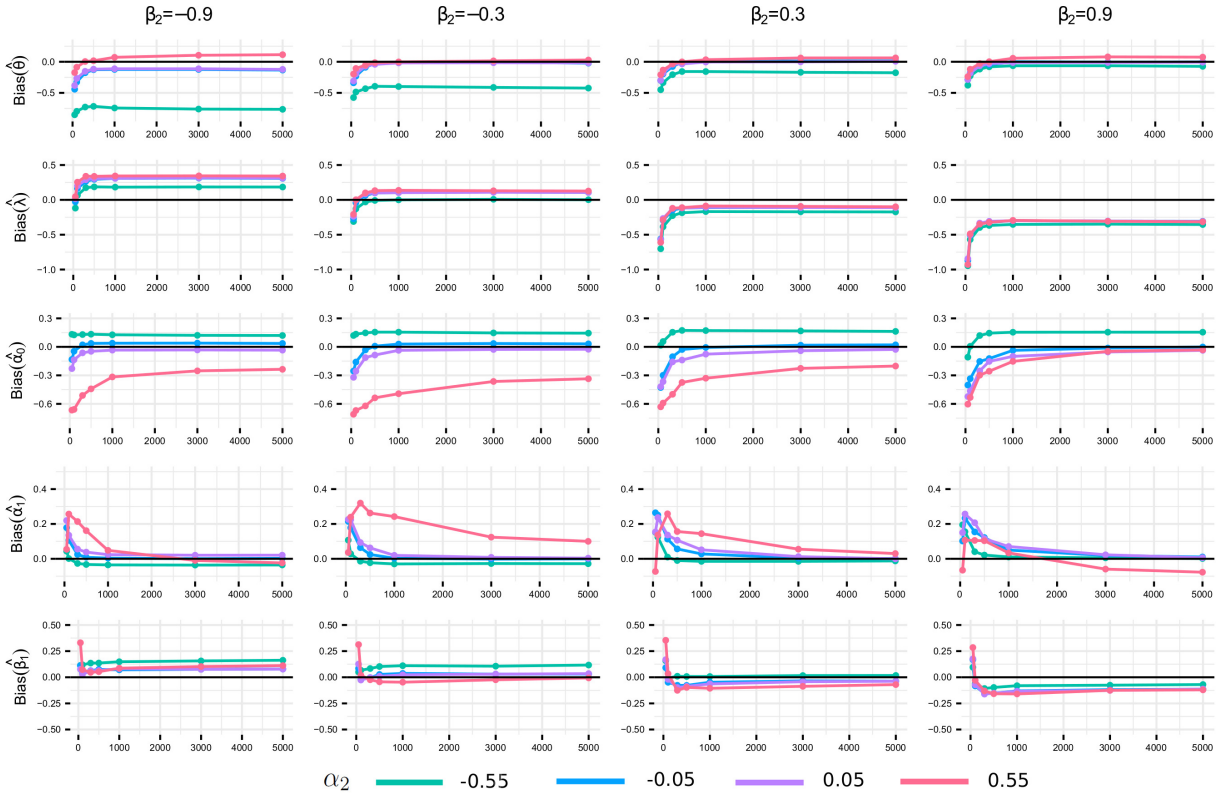

Figure 21 – Estimated parameter biases associated to fitted incomplete model considering different parameter α_2 and β_2 values.

Source: Elaborated by the author.

follow-up conducted until 2018. The diagnosis of malignant neoplasm of bronchus and lung (C34 - ICD-10 diagnosis code)[1] were included in the sample. All records were provided by the FOSP and it can be downloaded in <http://www.fosp.saude.sp.gov.br>. As mentioned by Andrade *et al.* (2012) these policies serve as an instrument for oncology hospitals to prepare their protocols and improve care practices.

In our study the event of interest was defined as death due to cancer. To identify the effects of the observed independent variables on hazard function, such as gender, age at diagnosis, clinical stage, surgery, radiotherapy, and chemotherapy as well as to capture the unobserved heterogeneity are the main goals.

Initially, we present a descriptive analysis of observed covariates in Table 23. According to Table 23, 19,657 (63.61%) patients were male, 11,090 (35.89%) patients were younger (< 60 years-old), and the most of patients were in the clinical stage III or IV 25971 (84,05%) patients. Regarding treatment, 5,899 (19.09%) patients underwent surgery, 11,709 (38.16%) patients received radiotherapy, and 20,134 (65.16%) patients received chemotherapy. A total of 27,479 (88.93%) events occurred during the follow-up period. The maximum observation time

[1] ICD-10 is the 10^{th} revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO).

was approximately 18.76 years, while the median follow-up time was 5.92 years.

Table 23 – Descriptive analysis of the observed covariates from the lung cancer dataset.

Covariate	Code	Category	Number of patients ($n = 30,900$)	%
X_1: Gender	0	Male	19,657	63.61%
	1	Female	11,243	36.39%
X_2: Age at diagnosis	0	Younger	11,090	35.89%
	1	Older	19,810	64.11%
X_3: Surgery	0	No	25,001	80.91%
	1	Yes	5,899	19.09%
X_4: Clinical Stage (CS)	$(1,0,0)$	I	3,058	9.90%
	$(0,1,0)$	II	1,871	6.06%
	$(0,0,1)$	III	9,200	29.77%
	$(0,0,0)$	IV	16,771	54.27%
X_5: Radiotherapy	0	No	19,109	61.84%
	1	Yes	11,791	38.16%
X_6: Chemotherapy	0	No	10,766	34.84%
	1	Yes	20,134	65.16%

Source: Elaborated by the author.

Figure 22 presents the overall estimated survival function obtained by KM estimator. The survival rate appears to trend reasonably close to 0 as the lifetime increased, as expected, once the most of patients were in the clinical stage III or IV. The median lung specific-survival period was approximately 0.726 years. The 0.5-, 1-, 2-, 5-, and 10-year specific survival rates are 0.599, 0.398, 0.215, 0.092, and 0.053, respectively.

Initially, we provide in Figure 23 a plot of log cumulative baseline hazards against time (follow-up period) for gender, age at diagnosis, surgery, clinical stage, radiotherapy and chemotherapy, respectively. According to Klein and Moeschberger (2003), if the proportionality assumption holds, then these curves should be approximately parallel, with constant vertical separation between them. The plots suggest that the hazard are non-proportional for the radiotherapy and chemotherapy covariates. For the surgery covariate the proportionality is questionable before 5 years, and is more evident for the radiotherapy and chemotherapy covariates, once that occurred a intersection between the curves.

The results of PH assumption testing for a Cox regression model fit (GRAMBSCH; THERNEAU, 1994) are displayed in Table 24; they provided strong evidence that the radiotherapy, chemotherapy and clinical stage variables have a non-constant effect over time, while the age at diagnosis and gender and surgery there are evidence of constant effect over time at 5% significance level. However, when a 10% significance level is considered there is also evidence of non-proportionality hazards for the surgery covariate.

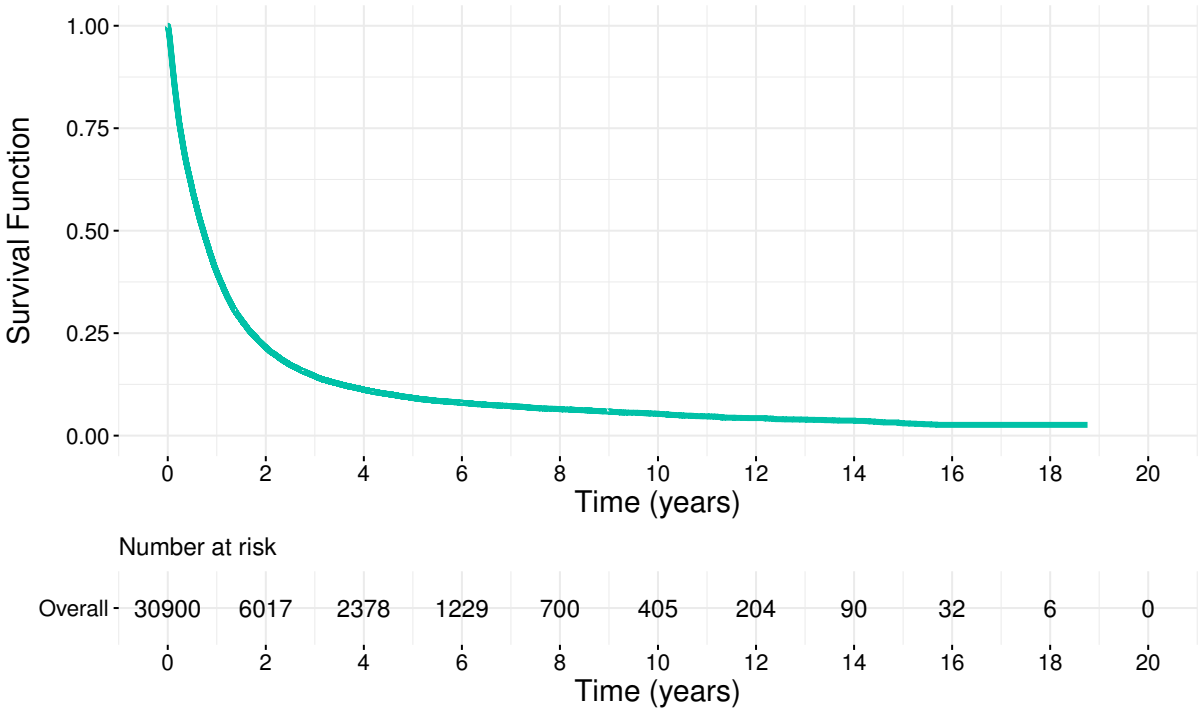

Figure 22 – Estimated survival curve obtained via KM for the lung cancer dataset.

Source: Elaborated by the author.

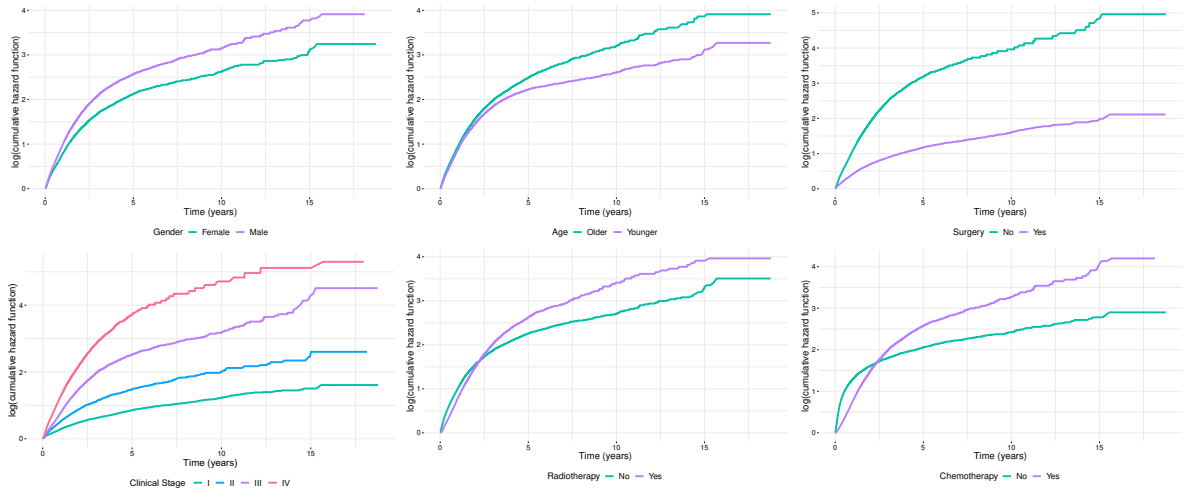

Figure 23 – Plot of log cumulative baseline hazard versus time of follow-up for the gender, age, surgery, clinical stage, radiotherapy and chemoterapy.

Source: Elaborated by the author.

To evaluate the effect of the observed covariates in the hazard function, as well as in the time effect , we fitted the traditional GTDL and GTDL-RWLF models to the dataset. For illustrative purposes, we link parameter α to covariates through an identity link function, as described in Equation (6.4). In this study, we adopted the same subset covariates on the two components ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$).

Table 24 – Test of PHs assumption.

Variable	ρ	χ^2	p-value
Gender	0.004	0.462	0.497
Age at diagnosis	−0.005	0.692	0.405
Surgery	−0.011	3.000	0.083
Clinical Stage II	−0.003	0.229	0.632
Clinical Stage III	0.024	15.857	<0.001
Clinical Stage IV	0.004	0.534	0.465
Radiotherapy	0.223	1374	<0.001
Chemotherapy	0.454	5643	<0.001

Source: Elaborated by the author.

The results of the fitted GTDL and GTDL-RWLF models are given in Table 25. According to the AIC criterion value, the GTDL-RWLF model seems to be the best choice between the GTDL model for all fitted models. Regardless of the fitted models, a significant effect in the lifetime for all observed covariates as the 95% confidence interval for β does not include 0. Besides, the time effect measure differs between groups (α_0 and α_1 are significant) for all covariates under GTDL model, while in the proposed model the time effects differ between groups for the gender, surgery, clinical stage III, radiotherapy and chemotherapy.

Considering the AIC criterion values, max $\ell(\cdot)$ values, and the number of parameters in the models, we select the GTDL-RWLF as our working model. We focused exclusively on an interpretation of GTDL-RWLF model parameters. Note that for surgery covariate the $\widehat{\alpha}_0 > 0$ and $\widehat{\alpha}_0 + \widehat{\alpha}_1 < 0$, which means that the distribution is proper in the no surgery group, while it is improper in the surgery group, leading to long-term survivors $\widehat{p}_1 = 0.138$; CI(95%) $= [0.118; 0.158]$. For the clinical stage, the estimated time effects was $\widehat{\alpha}_0 + \widehat{\alpha}_1 = -0.181$; CI(95%) $= [-0.205; -0.156]$ in the clinical stage I; $\widehat{\alpha}_0 + \widehat{\alpha}_2 = -0.219$; CI(95%) $= [-0.254; -0.183]$ in the clinical stage II; $\widehat{\alpha}_0 + \widehat{\alpha}_3 = -0.169$; CI(95%) $= [-0.198; -0.141]$ in the clinical stage III and $\widehat{\alpha}_0 = -0.216$; CI(95%) $= [-0.265; -0.166]$ in the clinical stage IV. As the time effect are negative, the model suggests that there are long-term survivors in each subgroup; the estimated proportions are, respectively, $\widehat{p}_{\text{I}} = 0.233$; CI(95%) $= [0.203; 0.262]$, $\widehat{p}_{\text{II}} = 0.104$; CI(95%) $= [0.081; 0.127]$, $\widehat{p}_{\text{III}} = 0.017$; CI(95%) $= [0.012; 0.022]$ and $\widehat{p}_{\text{IV}} = 0.005$; CI(95%) $= [0.003; 0.007]$. Although the estimates long-term survivors are close to zero (in each group), a better prognosis is associated with an early clinical stage. In addition, note that there is a significant difference in the long-term survivors among the clinical stages.

The other observed covariates such as gender, age, radiotherapy and chemotherapy the effect time was positive, which leads to the a proper survival function, that is, there is not evidence of long-term survivors. Note that the estimates of θ is greater than 0.6 for all fitted frailty models, except to clinical stage, which indicates a reasonable degree of unobserved heterogeneity in

the sample. Overall, the fitted models reasonably fit KM curves. However, the GTDL-RWLF model enables quantifying unobserved heterogeneity, which is of great importance in clinical practice, once those important covariates were not observed such as smoking, secondhand smoke, personal or family history of lung cancer, exposure to asbestos[2].

Figure 24 shows the estimated survival functions from the GTDL-RWLF model for each observed covariate. The survival function estimates are close to the KM curves. Note that the estimated curves for radiotherapy and chemotherapy covariates also cross over time. Such crossing can not occur in the traditional GTDL model (considering the effect time α equals in both groups) proposed by Mackenzie (1996), which is a disadvantage. The inclusion of a covariate through the α parameter allowed the quantification of each group of patients' effect, and yielded the curves to cross, as can be seen in estimated survival function (Figure 24).

We also consider the risk factors previously mentioned in a full model and the results of the fitted GTDL and GTDL-RWLF models are given in Table 26. According to the AIC criterion values, GTDL-RWLF model seem to be the best choice. Among the observed covariates considered in the models, there is evidence that all variables are important factors to explain the failure rate and the time effect because the 95% confidence interval of the coefficients $\boldsymbol{\beta}^\top = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8)$ and $\boldsymbol{\alpha}^\top = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8)$ do not include 0 regarding frailty model. Note that $\widehat{\theta} = 0.993$, which indicates a reasonable degree of unobserved heterogeneity in the sample.

The GTDL-RWLF model allows quantifying the amount of unobserved heterogeneity as previously mentioned. In this sense is important to test the suitability of the frailty term in the frailty model using theLR test given by, $\Lambda = 2\{\ell(\widehat{\Psi}) - \ell(\widehat{\Psi}_0)\}$, where $\widehat{\Psi}_0$ is the maximum likelihood estimator of $\widehat{\Psi}$ under the null hypothesis $H_0 : \theta = 0$. As the parameter value is on the boundary of the parametric space, Maller and Zhou (1996) showed that the statistic 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 $\Lambda = 1,055.28$ (p-value< 0.0001), which provides a strong evidence in favor of the inclusion of the frailty term.

According to the results, as expected, slightly better survival rates were associated with young patients, female in the clinical stage I and undergoing surgery. Before two years, a better survival rates were observed for patients who did radiotherapy/chemotherapy. After two years the effect of the treatment is lost and the survival rates are close in both treatments.

Table 27 shows the estimated survival rates at 0.5-, 1-, 2- and 10-year according to the GTDL-RWLF model for older patients and some combinations of covariates clinical stages, surgery, gender, radiotherapy and chemotherapy. As expected, patients with early-stage (clinical stage I), female patients who have undergone surgery and who did not radiotherapy and

[2] People who work with asbestos, such as in mills, mines, plants textile.

Table 25 – MLEs, SEs, 95% CIs, AIC value obtained for the traditional GTDL and GTDL-RWLF models considering gender, age at diagnosis, surgery, clinical stage, radiotherapy and chemotherapy fitted for the lung cancer dataset.

Model	GTDL model				GTDL-RWLF model			
			CI 95%				CI 95%	
Parameter	MLE	SE	Lower	Upper	MLE	SE	Lower	Upper
α_0	−0.467	0.007	−0.482	−0.452	1.537	0.127	1.288	1.786
$\alpha_{1\,(\text{Female})}$	0.059	0.013	0.034	0.083	−1.467	0.129	−1.719	−1.214
$\beta_{(\text{Female})}$	−0.430	0.024	−0.478	−0.383	−0.122	0.049	−0.218	−0.025
λ	2.197	0.020	2.157	2.236	2.380	0.045	2.292	2.468
θ	–				0.829	0.015	0.800	0.859
max $\ell(\cdot)$	−35,572.57				−35,293.64			
AIC	71,153.14				70,597.28			
α_0	−0.488	0.009	−0.506	−0.470	1.673	0.164	1.352	1.995
$\alpha_{1\,(\text{Older})}$	0.043	0.013	0.017	0.068	−0.208	0.187	−0.574	0.158
$\beta_{(\text{Older})}$	0.065	0.028	0.010	0.119	0.399	0.077	0.249	0.549
λ	1.957	0.024	1.910	2.003	2.006	0.050	1.907	2.104
θ					0.870	0.013	0.844	0.896
max $\ell(\cdot)$	−35,729.95				−35,379.72			
AIC	71,467.90				70,769.43			
α_0	−0.400	0.007	−0.415	−0.386	1.110	0.082	0.950	1.270
$\alpha_{1\,(\text{Surgery})}$	0.139	0.012	0.116	0.163	−1.242	0.082	−1.402	−1.081
$\beta_{(\text{Surgery})}$	−1.552	0.027	−1.605	−1.499	−1.443	0.034	−1.510	−1.376
λ	2.394	0.019	2.357	2.431	2.532	0.037	2.460	2.604
θ					0.626	0.011	0.605	0.647
max $\ell(\cdot)$	−33,605.09				−33,460.14			
AIC	67,218.17				66,930.28			
α_0	−0.428	0.011	−0.449	−0.407	−0.216	0.025	−0.265	−0.166
$\alpha_{1\,(\text{CS - I})}$	0.220	0.016	0.188	0.252	0.035	0.026	−0.016	0.086
$\alpha_{2\,(\text{CS - II})}$	0.155	0.020	0.116	0.194	−0.003	0.027	−0.056	0.050
$\alpha_{3\,(\text{CS - III})}$	0.150	0.014	0.122	0.177	0.046	0.019	0.008	0.085
$\beta_{(\text{CS - I})}$	−2.203	0.038	−2.278	−2.128	−2.249	0.040	−2.328	−2.171
$\beta_{(\text{CS - II})}$	−1.448	0.044	−1.535	−1.362	−1.475	0.046	−1.566	−1.384
$\beta_{(\text{CS - III})}$	−0.826	0.024	−0.872	−0.780	−0.851	0.026	−0.901	−0.801
λ	2.830	0.027	2.776	2.884	3.063	0.039	2.986	3.139
θ					0.205	0.020	0.166	0.244
max $\ell(\cdot)$	−32,869.38				−32,815.27			
AIC	65,754.76				65,648.54			
α_0	−0.575	0.008	−0.591	−0.559	−0.026	0.020	−0.066	0.014
$\alpha_{1\,(\text{Radiotherapy})}$	0.306	0.012	0.282	0.330	1.761	0.071	1.622	1.900
$\beta_{(\text{Radiotherapy})}$	−0.544	0.024	−0.590	−0.497	−1.563	0.037	−1.635	−1.490
λ	2.289	0.022	2.246	2.332	3.352	0.048	3.258	3.446
θ					0.889	0.015	0.859	0.919
max $\ell(\cdot)$	−35,447.21				−34,640.59			
AIC	70,902.41				69,291.18			
α_0	−0.967	0.015	−0.996	−0.939	−0.106	0.020	−0.145	−0.067
$\alpha_{1\,(\text{Chemotherapy})}$	0.733	0.016	0.702	0.765	2.071	0.047	1.978	2.163
$\beta_{(\text{Chemotherapy})}$	−1.210	0.021	−1.250	−1.169	−2.586	0.030	−2.645	−2.526
λ	3.562	0.047	3.470	3.654	6.977	0.138	6.706	7.248
θ					1.219	0.014	1.191	1.247
max $\ell(\cdot)$	−34,344.24				−32,433.39			
AIC	68,696.48				64,876.77			

Source: Elaborated by the author.

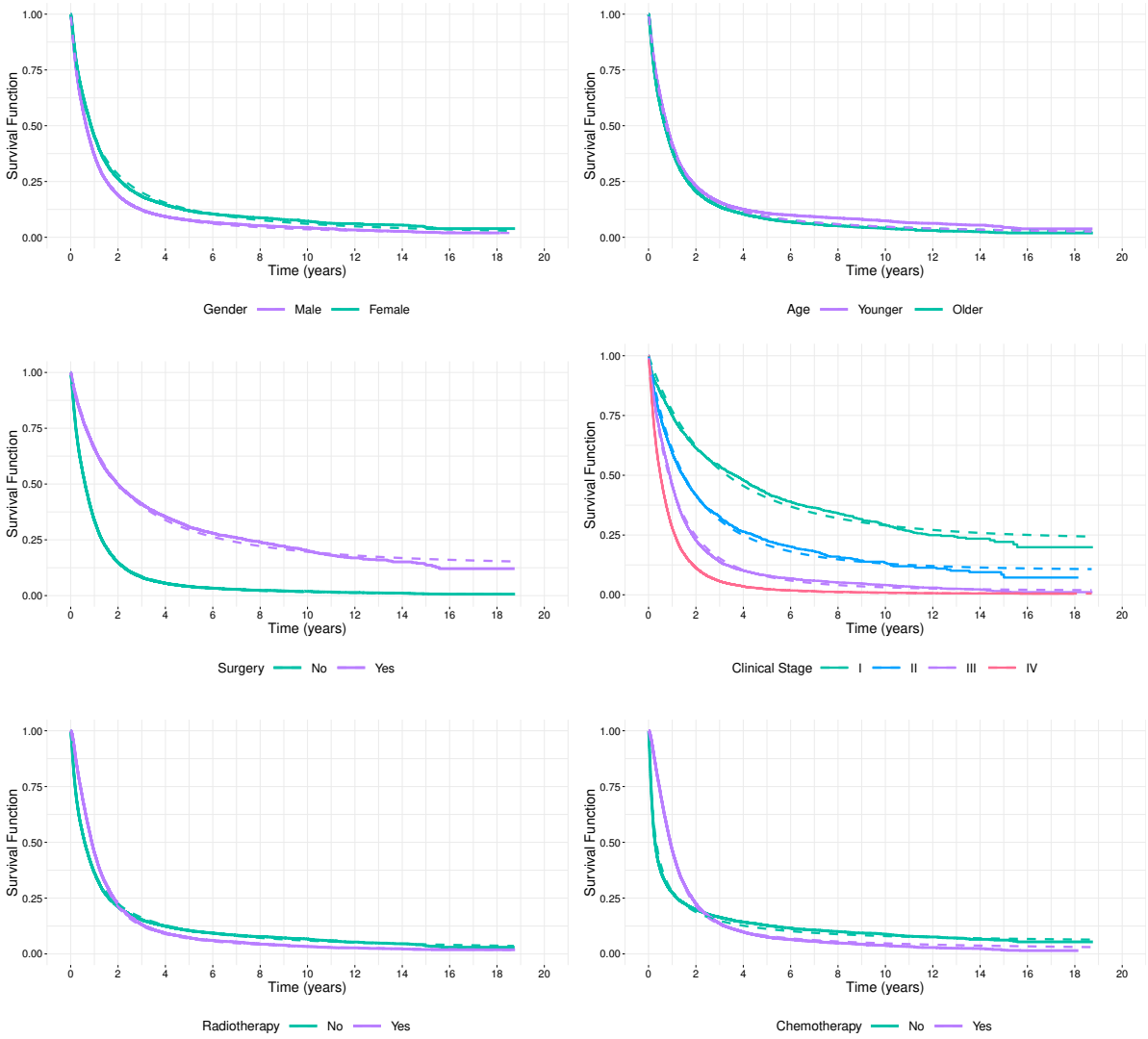

Figure 24 – Estimated survival curve obtained via KM (full line) for lung cancer dataset, and estimated survival function according to GTDL-RWLF model (dashed line) for gender, age at diagnosis, surgery, clinical stage, radiotherapy and chemotherapy.

Source: Elaborated by the author.

chemotherapy have higher survival at 10-year; whereas patients diagnosed in clinical stage IV have worst prognosis, regardless of the gender, treatment received and the estimated survival rates at 10-year are close to zero for these patients. Overall the estimated survival rates at 0.5-year are highly impacted by clinical stage (late diagnosis), once that the most of patients were with metastatic disease (clinical stage IV).

6.5 Concluding remarks

In this chapter, we have extended the GTDL model by including a RWL frailty term in the hazard function in order to quantify the amount of unobserved heterogeneity among

Table 26 – MLEs, SEs, 95% asymptotic CIs, AIC value obtained for the traditional GTDL and GTDL-RWLF models considering gender, age at diagnosis, surgery, clinical stage, radiotherapy and chemotherapy fitted for the lung dataset.

	GTDL model				GTDL-RWLF model			
			CI 95%				CI 95%	
Parameter	MLE	SE	Lower	Upper	MLE	SE	Lower	Upper
α_0	-0.794	0.017	-0.827	-0.761	1.445	0.075	1.298	1.591
$\alpha_{1\,(\text{Female})}$	0.046	0.009	0.027	0.064	-0.138	0.025	-0.186	-0.089
$\alpha_{2\,(\text{Older})}$	0.092	0.009	0.074	0.111	0.198	0.025	0.150	0.247
$\alpha_{3\,(\text{Surgery})}$	0.111	0.011	0.090	0.133	-0.628	0.054	-0.734	-0.521
$\alpha_{4\,(\text{CS - I})}$	0.226	0.015	0.197	0.255	-1.108	0.067	-1.239	-0.976
$\alpha_{5\,(\text{CS - II})}$	0.074	0.017	0.041	0.107	-1.246	0.068	-1.380	-1.112
$\alpha_{6\,(\text{CS - III})}$	0.020	0.011	-0.003	0.042	-0.929	0.061	-1.048	-0.809
$\alpha_{7\,(\text{Radiotherapy})}$	0.139	0.010	0.120	0.157	0.333	0.033	0.268	0.398
$\alpha_{8\,(\text{Chemotherapy})}$	0.470	0.013	0.445	0.496	0.484	0.026	0.432	0.536
$\beta_{1\,(\text{Female})}$	-0.305	0.018	-0.341	-0.269	-0.299	0.028	-0.354	-0.244
$\beta_{2\,(\text{Older})}$	-0.075	0.018	-0.111	-0.039	-0.141	0.028	-0.196	-0.086
$\beta_{3\,(\text{Surgery})}$	-1.004	0.026	-1.056	-0.953	-1.027	0.042	-1.109	-0.944
$\beta_{4\,(\text{CS - I})}$	-1.933	0.038	-2.008	-1.858	-2.009	0.057	-2.120	-1.898
$\beta_{5\,(\text{CS - II})}$	-1.024	0.040	-1.103	-0.946	-0.747	0.061	-0.866	-0.628
$\beta_{6\,(\text{CS - III})}$	-0.467	0.020	-0.506	-0.428	-0.311	0.034	-0.379	-0.244
$\beta_{7\,(\text{Radiotherapy})}$	-0.440	0.018	-0.476	-0.405	-0.864	0.029	-0.922	-0.806
$\beta_{8\,(\text{Chemotherapy})}$	-1.428	0.019	-1.465	-1.392	-2.103	0.029	-2.160	-2.047
λ	8.049	0.138	7.779	8.320	11.239	0.295	10.661	11.818
θ	–	–	–	–	0.993	0.016	0.962	1.023
$\max \ell(\cdot)$	$-29,340.89$				$-28,813.25$			
AIC	58,717.78				57,664.49			

Source: Elaborated by the author.

individuals under study. An advantage of the GTDL-RWLF model over alternatives is that it does not make assumptions about the existence of long-term survivors or subgroups with evidence of long-term survivors since the survival function can be proper or improper; this makes the model flexible and applicable to situations with or without (or both) long-term survivors. Besides, it captures different time effects in the subgroups, which is biologically plausible. Simulation studies showed that MLEs' frequentist properties (consistency and asymptotic normality) of the GTDL-RWLF model parameters were satisfied when the sample size increases, as expected. Also, we have verified through simulation studies that the GTDL-RWLF model can identify the accurate way the existence or not of long-term survivors in a subgroup and PH when, in fact, there is the proportionally of hazards. Besides, misspecification simulation studies showed that the GDTL-RWLF model can still yield reasonable estimates. The GTDL-RWL model's practical relevance and applicability were demonstrated using a real and novel lung cancer dataset, with

Table 27 – Estimated specific survival rates at 0.5-, 1-, 2- and 10-year for the GTDL-RWLF model considering clinical stage, surgery, gender, radiotherapy and chemotherapy for older patients.

Clinical stage	Surgery	Gender	Radiotherapy	Chemotherapy	$S(0.5)$	$S(1)$	$S(2)$	$S(10)$
I	No	Female	No	No	0.672	0.485	0.287	0.031
				Yes	0.932	0.843	0.616	0.028
			Yes	No	0.809	0.638	0.378	0.025
				Yes	0.967	0.912	0.707	0.025
		Male	No	No	0.603	0.406	0.220	0.024
				Yes	0.907	0.787	0.506	0.024
			Yes	No	0.755	0.554	0.290	0.022
				Yes	0.954	0.877	0.602	0.023
	Yes	Female	No	No	0.862	0.768	0.649	0.432
				Yes	0.978	0.954	0.901	0.351
			Yes	No	0.931	0.868	0.757	0.294
				Yes	0.990	0.976	0.936	0.120
		Male	No	No	0.820	0.699	0.550	0.262
				Yes	0.970	0.935	0.853	0.150
			Yes	No	0.907	0.820	0.669	0.133
				Yes	0.986	0.966	0.903	0.058
IV	No	Female	No	No	0.279	0.149	0.073	0.016
				Yes	0.603	0.327	0.117	0.017
			Yes	No	0.376	0.192	0.084	0.017
				Yes	0.753	0.449	0.139	0.018
		Male	No	No	0.250	0.135	0.068	0.016
				Yes	0.534	0.272	0.101	0.017
			Yes	No	0.329	0.167	0.076	0.016
				Yes	0.690	0.374	0.118	0.017
	Yes	Female	No	No	0.447	0.261	0.124	0.018
				Yes	0.821	0.614	0.279	0.020
			Yes	No	0.604	0.369	0.158	0.018
				Yes	0.907	0.748	0.350	0.020
		Male	No	No	0.389	0.218	0.103	0.017
				Yes	0.769	0.529	0.214	0.019
			Yes	No	0.534	0.305	0.127	0.018
				Yes	0.875	0.673	0.268	0.019

Source: Elaborated by the author.

characteristics of NPHs, long-term survivors, and the intersection of survival functions for some variables. In this real example, we concluded through AIC that the fit of the GTDL-RWLF model was better than the traditional GTDL model. As a finding, we reported that, as expected, slightly better survival rates were associated with patients who are young, female, in clinical stage I, and underwent surgery. Before two years, better survival rates were observed for patients who received radiotherapy or chemotherapy. After two years, the treatment's effect was lost, and both treatments' survival rates were similar.

CHAPTER

7

FINAL REMARKS

Survival data analysis has played an important role in diverse areas of knowledge. In this work, we have proposed different methodologies for modeling survival data based on the RWL distribution. Under this parameterization, one of the parameters is given by the mean, whereas the other parameter can be interpreted as a precision parameter. We found the moments, harmonic mean, mean and median deviations, and Laplace transform of this distribution. The use of the ML estimation under random censoring was discussed in detail. In the real proposed applications, we observed that the RWL distribution returned a better fit when compared to similar reparameterized distributions.

The Laplace transform of the RWL distribution provided an easy mathematical treatment for obtaining analytical expressions for unconditional survival and hazard functions of the RWL frailty model. We showed that the RWL frailty model with Weibull returned better results in the simulation studies than with the baseline Gompertz model. In the application with lung cancer data, the RWL frailty models showed to be useful to capture the unobserved heterogeneity and were highly competitive in terms of fitting when compared with gamma, BS, and IG frailty models. We reported that the Cox PH model had the worst fit for the data. According to the RWL frailty model with Weibull baseline, we concluded that female patients exhibited slightly better survival than male patients with the same clinical stage and treatment, as expected. Meanwhile, as also expected, the survival rates are worse in the absence of treatment, mainly in the later years of diagnosis.

We have introduced the unified long-term model with a RWL frailty term for modeling jointly cure fraction and the frailty among non-cured patients. In this approach, the idea is that the observation of an event of interest (e.g., the patient's death) may be due to one or some competing causes. The number of latent competing causes was assumed to follow a binomial distribution, which is flexible and accommodates underdispersion, equidispersion, and overdispersion. We included the cured fraction as a model parameter and linked it to covariates. Considering the ML method, the PoCrRWLF model presented great results in the simulated study. In the application

with stomach cancer data, the model was useful to handle the cure fraction and quantify the unobserved frailty among non-cured patients.

In modeling of survival data with proportional and non-proportional hazards, we proposed the GTDL-RWLF model, which is an extension of GTDL model including a RWL frailty term to quantify the amount of unobserved heterogeneity among individuals under study. This model is flexible and applicable to situations with or without (or both) long-term survivors. Besides, it captures different time effects in the subgroups, which is biologically plausible. Simulation studies showed that MLEs' asymptotic properties of the GTDL-RWLF model parameters were satisfied when the sample size increases, as expected. Also, we have verified through simulation studies that the GTDL-RWLF model can identify the accurate way the existence or not of long-term survivors in a subgroup and PH when, in fact, there is the proportionally of hazards. Moreover, misspecification simulation studies showed that the GDTL-RWLF model can still yield reasonable estimates. An application to lung cancer was presented to illustrate its modeling capability. In this real example, the GTDL-RWLF model was better than the traditional GTDL model. According to GTDL-RWLF model, we concluded slightly better survival rates were associated with patients who are young, female, in clinical stage I, and underwent surgery. Before two years, better survival rates were observed for patients who received radiotherapy or chemotherapy. After two years, the treatment's effect was lost, and both treatments' survival rates were similar.

Future research

There are many possible extensions of the current work to consider further. Some are given as follows.

- To develop other estimation methods for estimating the parameters of RWL frailty model
- To derive influence diagnostic tools in the regression cases to evaluate the effect of atypical observations on the model
- To propose the accelerated failure time frailty model for modeling of multiple systems subject to minimal repair

BIBLIOGRAPHY

AALEN, O. Nonparametric inference for a family of counting processes. **The Annals of Statistics**, JSTOR, p. 701–726, 1978. Citation on page 31.

____. A model for nonparametric regression analysis of counting processes. In: **Mathematical statistics and probability theory**. [S.l.]: Springer, 1980. p. 1–25. Citations on pages 23 and 35.

AALEN, O. O. Heterogeneity in survival analysis. **Statistics in medicine**, Wiley Online Library, v. 7, n. 11, p. 1121–1137, 1988. Citations on pages 25 and 26.

AFIFY, A. Z.; NASSAR, M.; CORDEIRO, G. M.; KUMAR, D. The weibull marshall–olkin lindley distribution: properties and estimation. **Journal of Taibah University for Science**, Taylor & Francis, v. 14, n. 1, p. 192–204, 2020. Citation on page 22.

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974. Citations on pages 54, 88, and 105.

AL-MUTAIRI, D.; GHITANY, M.; KUNDU, D. Inferences on stress-strength reliability from weighted lindley distributions. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 44, n. 19, p. 4096–4113, 2015. Citation on page 22.

AL-ZAHRANI, B.; GINDWAN, M. Parameter estimation of a two-parameter lindley distribution under hybrid censoring. **International Journal of System Assurance Engineering and Management**, Springer, v. 5, n. 4, p. 628–636, 2014. Citation on page 23.

ALI, S. On the bayesian estimation of the weighted lindley distribution. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 85, n. 5, p. 855–880, 2015. Citations on pages 22 and 23.

ALMEIDA, M. P.; PAIXÃO, R. S.; RAMOS, P. L.; TOMAZELLA, V.; LOUZADA, F.; EHLERS, R. S. Bayesian non-parametric frailty model for dependent competing risks in a repairable systems framework. **Reliability Engineering & System Safety**, Elsevier, v. 204, p. 107145, 2020. Citation on page 25.

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. Citations on pages 73 and 106.

ARANDA-ORDAZ, F. J. An extension of the proportional-hazards model for grouped data. **Biometrics**, JSTOR, p. 109–117, 1983. Citation on page 23.

ASGHARZADEH, A.; BAKOUCH, H. S.; NADARAJAH, S.; SHARAFI, F. *et al.* A new weighted lindley distribution with application. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 30, n. 1, p. 1–27, 2016. Citation on page 23.

ASGHARZADEH, A.; NADARAJAH, S.; SHARAFI, F. Weibull lindley distribution. **REVSTAT Statistical Journal**, v. 16, p. 87–113, 2018. Citation on page 21.

BAKOUCH, H. S.; AL-ZAHRANI, B. M.; AL-SHOMRANI, A. A.; MARCHI, V. A.; LOUZADA, F. An extended lindley distribution. **Journal of the Korean Statistical Society**, Elsevier, v. 41, n. 1, p. 75–85, 2012. Citation on page 21.

BALAKRISHNAN, N.; PENG, Y. Generalized gamma frailty model. **Statistics in medicine**, Wiley Online Library, v. 25, n. 16, p. 2797–2816, 2006. Citations on pages 26 and 40.

BARKER, P.; HENDERSON, R. Small sample bias in the gamma frailty model for univariate survival. **Lifetime data analysis**, Springer, v. 11, n. 2, p. 265–284, 2005. Citations on pages 38, 39, and 69.

BARRIGA, G. D.; CANCHO, V. G.; GARIBAY, D. V.; CORDEIRO, G. M.; ORTEGA, E. M. A new survival model with surviving fraction: An application to colorectal cancer data. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 28, n. 9, p. 2665–2680, 2019. Citation on page 99.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, Taylor & Francis, v. 47, n. 259, p. 501–515, 1952. Citations on pages 22, 24, 25, 38, and 82.

BILLINGSLEY, P. **Probability and measure**. [S.l.]: John Wiley & Sons, 2008. Citation on page 46.

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 24.

BÖHNSTEDT, M.; GAMPE, J.; PUTTER, H. Information measures and design issues in the study of mortality deceleration: findings for the gamma-gompertz model. **Lifetime Data Analysis**, Springer, p. 1–24, 2021. Citation on page 62.

BORUCKA, J. Extensions of cox model for non-proportional hazards purpose. **Ekonometria**, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, n. 45, p. 85–101, 2014. Citation on page 23.

BOURGUIGNON, M. An alternative conjugate prior distribution for positive parameters. **Annals of Data Science**, Springer, v. 6, n. 2, p. 237–243, 2019. Citation on page 22.

BOURGUIGNON, M.; GALLARDO, D. I. Reparameterized inverse gamma regression models with varying precision. **Statistica Neerlandica**, Wiley Online Library, 2020. Citations on pages 26 and 54.

BOURGUIGNON, M.; SANTOS-NETO, M.; CASTRO, M. de. A new regression model for positive random variables with skewed and long tail. **Metron**, Springer, v. 79, n. 1, p. 33–55, 2021. Citation on page 26.

BOX-STEFFENSMEIER, J. M.; ZORN, C. J. Duration models and proportional hazards in political science. **American Journal of Political Science**, JSTOR, p. 972–988, 2001. Citation on page 23.

BOZDOGAN, H. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. **Psychometrika**, Springer, v. 52, n. 3, p. 345–370, 1987. Citation on page 54.

BRENT, R. P. **Algorithms for Minimization without Derivatives, chap. 4**. [S.l.]: Prentice-Hall Englewood Cliffs, NJ, USA, 1973. Citation on page 45.

BRETAGNOLLE, J.; HUBER-CAROL, C. Effects of omitting covariates in cox's model for survival data. **Scandinavian journal of statistics**, JSTOR, p. 125–138, 1988. Citation on page 25.

BRYSON, M. C.; SIDDIQUI, M. Some criteria for aging. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 64, n. 328, p. 1472–1483, 1969. Citation on page 48.

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**, p. 1–19, 2019. Citations on pages 24, 35, 36, 96, and 97.

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**, Wiley Online Library, v. 61, p. 841–859, 2019. Citation on page 63.

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**, Taylor & Francis, v. 46, n. 13, p. 2434–2459, 2019. Citation on page 63.

CANCHO, V. G.; MACERA, M. A.; SUZUKI, A. K.; LOUZADA, F.; ZAVALETA, K. E. A new long-term survival model with dispersion induced by discrete frailty. **Lifetime data analysis**, Springer, v. 26, n. 2, p. 221–244, 2020. Citation on page 25.

CASTRO, M. d.; CANCHO, V. G.; RODRIGUES, J. A bayesian long-term survival model parametrized in the cured fraction. **Biometrical Journal: Journal of Mathematical Methods in Biosciences**, Wiley Online Library, v. 51, n. 3, p. 443–455, 2009. Citations on pages 37, 80, and 84.

CEPEDA, E.; GAMERMAN, D. Bayesian methodology for modeling parameters in the two parameter exponential family. **Revista Estadística**, v. 57, n. 168-169, p. 93–105, 2005. Citation on page 26.

CHEN, M.-H.; IBRAHIM, J. G.; SINHA, D. A new bayesian model for survival data with a surviving fraction. **Journal of the American Statistical Association**, Taylor & Francis, v. 94, n. 447, p. 909–919, 1999. Citation on page 25.

CHEN, Z. A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. **Statistics & Probability Letters**, Elsevier, v. 49, n. 2, p. 155–161, 2000. Citation on page 22.

COLOSIMO, E.; GIOLO, S. Análise de sobrevivência aplicada. 1ª edição. **São Paulo: Editora Edgard Blücher**, 2006. Citations on pages 21, 23, 30, 31, 35, and 50.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. Citations on pages 23 and 34.

COX, D. R. Partial likelihood. **Biometrika**, Oxford University Press, v. 62, n. 2, p. 269–276, 1975. Citation on page 35.

COX, D. R.; REID, N. Parameter orthogonality and approximate conditional inference. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 49, n. 1, p. 1–18, 1987. Citation on page 26.

COX, D. R.; SNELL, E. J. A general definition of residuals. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 30, n. 2, p. 248–265, 1968. Citation on page 75.

DANIEL, W. W. Applied nonparametric statistics. PWS-Kent Pub. Boston, 1990. Citation on page 55.

DAVID, G. K.; MITCHEL, K. **Survival Analysis: A Self-Learning Text**. [S.l.]: Spinger, 2012. Citation on page 29.

DUCHATEAU, L.; JANSSEN, P. **The frailty model**. [S.l.]: Springer Science & Business Media, 2007. Citation on page 25.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citations on pages 89 and 91.

ELBERS, C.; RIDDER, G. True and spurious duration dependence: The identifiability of the proportional hazard model. **The Review of Economic Studies**, Wiley-Blackwell, v. 49, n. 3, p. 403–409, 1982. Citations on pages 39 and 62.

ETEZADI-AMOLI, J.; CIAMPI, A. Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function. **Biometrics**, JSTOR, p. 181–192, 1987. Citations on pages 23 and 35.

EVERT, S.; BARONI, M.; EVERT, M. S. The zipfR package. **Available on-line at URL: http://cran. r-project. org/doc/packages/zipfR. pdf**, 2006. Citation on page 43.

FELLER, W. **An introduction to probability theory and its applications**. [S.l.]: John Wiley & Sons, 2008. Citations on pages 25 and 37.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of applied statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004. Citation on page 26.

FISHER, L. D.; LIN, D. Y. Time-dependent covariates in the cox proportional-hazards regression model. **Annual review of public health**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 20, n. 1, p. 145–157, 1999. Citation on page 23.

GALLARDO, D. I.; CASTRO, M. de; GÓMEZ, H. W. An alternative promotion time cure model with overdispersed number of competing causes: An application to melanoma data. **Mathematics**, Multidisciplinary Digital Publishing Institute, v. 9, n. 15, p. 1815, 2021. Citations on pages 25 and 38.

GALLARDO, D. I.; GÓMEZ-DÉNIZ, E.; LEÃO, J.; GÓMEZ, H. W. Estimation and diagnostic tools in reparameterized slashed rayleigh regression model. an application to chemical data. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, p. 104189, 2020. Citation on page 26.

GALLARDO, D. I.; GÓMEZ, H. W.; BOLFARINE, H. A new cure rate model based on the yule–simon distribution with application to a melanoma data set. **Journal of Applied Statistics**, Taylor & Francis, v. 44, n. 7, p. 1153–1164, 2017. Citations on pages 25 and 38.

GALLARDO, D. I.; GÓMEZ, Y. M.; CASTRO, M. de. A flexible cure rate model based on the polylogarithm distribution. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 88, n. 11, p. 2137–2149, 2018. Citation on page 25.

GALLARDO, D. I.; GOMEZ, Y. M.; GOMEZ, H. W.; CASTRO, M. de. On the use of the modified power series family of distributions in a cure rate model context. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 29, n. 7, p. 1831–1845, 2020. Citation on page 25.

GHITANY, M.; AL-MUTAIRI, D. K.; BALAKRISHNAN, N.; AL-ENEZI, L. Power lindley distribution and associated inference. **Computational Statistics & Data Analysis**, Elsevier, v. 64, p. 20–33, 2013. Citation on page 21.

GHITANY, M.; ALQALLAF, F.; AL-MUTAIRI, D. K.; HUSAIN, H. A two-parameter weighted lindley distribution and its applications to survival data. **Mathematics and Computers in simulation**, Elsevier, v. 81, n. 6, p. 1190–1201, 2011. Citations on pages 21, 22, 32, 33, 34, and 44.

GHITANY, M.; SONG, P.; WANG, S. New modified moment estimators for the two-parameter weighted lindley distribution. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 87, n. 16, p. 3225–3240, 2017. Citation on page 23.

GHITANY, M.; WANG, S. A note on parameter asymptotics for weighted lindley distribution. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, p. 1–12, 2019. Citations on pages 22 and 33.

GHITANY, M. E.; ATIEH, B.; NADARAJAH, S. Lindley distribution and its application. **Mathematics and computers in simulation**, Elsevier, v. 78, n. 4, p. 493–506, 2008. Citation on page 21.

GRAMBSCH, P. M.; THERNEAU, T. M. Proportional hazards tests and diagnostics based on weighted residuals. **Biometrika**, Biometrika Trust, v. 81, p. 515–526, 1994. Citation on page 107.

HA, I. D.; MACKENZIE, G. Robust frailty modelling using non-proportional hazards models. **Statistical modelling**, SAGE Publications Sage India: New Delhi, India, v. 10, n. 3, p. 315–332, 2010. Citations on pages 24 and 99.

HANNAN, E. J.; QUINN, B. G. The determination of the order of an autoregression. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 41, n. 2, p. 190–195, 1979. Citation on page 54.

HASNA, M. O.; ALOUINI, M.-S. Harmonic mean and end-to-end performance of transmission systems with relays. **IEEE Transactions on communications**, IEEE, v. 52, n. 1, p. 130–135, 2004. Citation on page 46.

HASTIE, T.; TIBSHIRANI, R. Varying-coefficient models. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 55, n. 4, p. 757–779, 1993. Citation on page 23.

HENDERSON, R.; OMAN, P. Effect of frailty on marginal regression estimates in survival analysis. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 61, n. 2, p. 367–379, 1999. Citations on pages 25 and 34.

HENNINGSEN, A.; TOOMET, O. maxlik: A package for maximum likelihood estimation in r. **Computational Statistics**, Springer, v. 26, n. 3, p. 443–458, 2011. Citations on pages 51 and 66.

HESS, K. R. Graphical methods for assessing violations of the proportional hazards assumption in cox regression. **Statistics in medicine**, Wiley Online Library, v. 14, n. 15, p. 1707–1723, 1995. Citation on page 35.

HOROWITZ, J. L. Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. **Econometrica**, Wiley Online Library, v. 67, n. 5, p. 1001–1028, 1999. Citation on page 25.

HOSMER, D. W.; LEMESHOW, S. **Applied survival analysis: Time-to-event**. [S.l.]: Wiley-Interscience, 1999. Citation on page 23.

HOUGAARD, P. Survival models for heterogeneous populations derived from stable distributions. **Biometrika**, Oxford University Press, v. 73, n. 2, p. 387–396, 1986. Citation on page 26.

____. Modelling heterogeneity in survival data. **Journal of Applied Probability**, JSTOR, p. 695–701, 1991. Citation on page 25.

____. Frailty models for survival data. **Lifetime data analysis**, Springer, v. 1, n. 3, p. 255–273, 1995. Citation on page 26.

____. **Analysis of multivariate survival data**. [S.l.]: Springer Science & Business Media, 2012. Citations on pages 25, 26, and 40.

IBRAHIM, J. G.; CHEN, M.-H.; SINHA, D. Cure rate models. In: **Bayesian Survival Analysis**. [S.l.]: Springer, 2001. p. 155–207. Citations on pages 24, 25, and 37.

INGLE, V.; KOGON, S.; MANOLAKIS, D. **Statisical and Adaptive Signal Processing**. [S.l.]: Artech, 2005. Citation on page 46.

KALBFLEISCH, J. D.; PRENTICE, R. L. **The statistical analysis of failure time data**. [S.l.]: John Wiley & Sons, 2011. Citation on page 35.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, v. 53, p. 457–481, 1958. Citation on page 88.

KASS, R. E.; RAFTERY, A. E. Bayes factors. **Journal of the american statistical association**, Taylor & Francis, v. 90, n. 430, p. 773–795, 1995. Citations on pages 73 and 75.

KEMALOGLU, S. A.; YILMAZ, M. Transmuted two-parameter lindley distribution. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 46, n. 23, p. 11866–11879, 2017. Citation on page 21.

KHAN, S. A. Exponentiated weibull regression for time-to-event data. **Lifetime data analysis**, Springer, v. 24, n. 2, p. 328–354, 2018. Citation on page 55.

KIM, H.-M.; JANG, Y.-H. New closed-form estimators for weighted lindley distribution. **Journal of the Korean Statistical Society**, Springer, v. 50, n. 2, p. 580–606, 2021. Citation on page 23.

KLEIN, J. P. Semiparametric estimation of random effects using the cox model based on the em algorithm. **Biometrics**, JSTOR, p. 795–806, 1992. Citation on page 39.

KLEIN, J. P.; MOESCHBERGER, M. L. Survival analysis: Statistical methods for censored and truncated data. **Springer Verlag, New York**, 2003. Citations on pages 24 and 107.

____. **Survival analysis: techniques for censored and truncated data**. [S.l.]: Springer Science & Business Media, 2006. Citations on pages 29 and 35.

KLEINBAUM, D. G.; KLEIN, M. Extension of the cox proportional hazards model for time-dependent variables. In: **Survival analysis**. [S.l.]: Springer, 2012. p. 241–288. Citation on page 23.

____. The stratified cox procedure. In: **Survival Analysis**. [S.l.]: Springer, 2012. p. 201–240. Citation on page 23.

LAWLESS, J. F. **Statistical models and methods for lifetime data**. [S.l.]: John Wiley & Sons, 2011. Citations on pages 21, 55, and 75.

LEÃO, J.; BOURGUIGNON, M.; GALLARDO, D. I.; ROCHA, R.; TOMAZELLA, V. A new cure rate model with flexible competing causes with applications to melanoma and transplantation data. **Statistics in Medicine**, Wiley Online Library, v. 39, n. 24, p. 3272–3284, 2020. Citations on pages 25, 37, and 38.

LEÃO, J.; LEIVA, V.; SAULO, H.; TOMAZELLA, V. Birnbaum–saunders frailty regression models: Diagnostics and application to medical data. **Biometrical Journal**, Wiley Online Library, v. 59, n. 2, p. 291–314, 2017. Citations on pages 26 and 73.

LEÃO, J.; LEIVA, V.; SAULO, H.; TOMAZELLA, V. *et al.* A survival model with birnbaum–saunders frailty for uncensored and censored cancer data. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 32, n. 4, p. 707–729, 2018. Citations on pages 26, 37, 84, and 91.

LEE, E. T.; WANG, J. **Statistical methods for survival data analysis**. [S.l.]: John Wiley & Sons, 2003. Citation on page 30.

LEHMANN, E. L. **Elements of large-sample theory**. [S.l.]: Springer Science & Business Media, 2004. Citations on pages 67 and 99.

LEHMANN, E. L.; CASELLA, G. **Theory of point estimation**. [S.l.]: Springer Science & Business Media, 2006. Citations on pages 65, 85, 98, and 99.

LIMBRUNNER, J. F.; VOGEL, R. M.; BROWN, L. C. Estimation of harmonic mean of a lognormal variable. **Journal of hydrologic engineering**, American Society of Civil Engineers, v. 5, n. 1, p. 59–66, 2000. Citation on page 46.

LIN, R. S.; LIN, J.; ROYCHOUDHURY, S.; ANDERSON, K. M.; HU, T.; HUANG, B.; LEON, L. F.; LIAO, J. J.; LIU, R.; LUO, X. *et al.* Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis. **Statistics in Biopharmaceutical Research**, Taylor & Francis, v. 12, n. 2, p. 187–198, 2020. Citation on page 23.

LINDLEY, D. V. Fiducial distributions and bayes' theorem. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 102–107, 1958. Citations on pages 21, 32, and 33.

LOUZADA, F.; CUMINATO, J. A.; RODRIGUEZ, O. M. H.; TOMAZELLA, V. L.; MILANI, E. A.; FERREIRA, P. H.; RAMOS, P. L.; BOCHIO, G.; PERISSINI, I. C.; JUNIOR, O. A. G. *et al.* Incorporation of frailties into a non-proportional hazard regression model and its diagnostics for reliability modeling of downhole safety valves. **IEEE Access**, IEEE, v. 8, p. 219757–219774, 2020. Citations on pages 26 and 91.

LOUZADA, F.; RAMOS, P. L. A new long-term survival distribution. **Biostatistics and Biometrics Open Access Journal**, Juniper Publishers Inc., v. 1, n. 5, p. 104–109, 2017. Citation on page 22.

LOUZADA-NETO, F. Extended hazard regression model for reliability and survival analysis. **Lifetime Data Analysis**, Springer, v. 3, n. 4, p. 367–381, 1997. Citation on page 24.

LOUZADA-NETO, F. Polyhazard models for lifetime data. **Biometrics**, Wiley Online Library, v. 55, n. 4, p. 1281–1285, 1999. Citation on page 24.

LOUZADA-NETO, F.; CREMASCO, C. P.; MACKENZIE, G. Sampling-based inference for the generalized time-dependent logistic hazard model. Gowas Publishers, 2010. Citation on page 24.

LUKACS, E. A survey of the theory of characteristic functions. **Advances in Applied Probability**, Cambridge University Press, v. 4, n. 1, p. 1–37, 1972. Citation on page 46.

MACKENZIE, G. Regression models for survival data: the generalized time-dependent logistic family. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 45, n. 1, p. 21–34, 1996. Citations on pages 24, 35, 36, and 110.

MALLER, R. A.; ZHOU, X. **Survival analysis with long-term survivors**. [S.l.]: John Wiley & Sons, 1996. Citations on pages 24, 25, 36, 63, 69, 75, 91, and 110.

MAZUCHELI, J.; COELHO-BARROS, E. A.; ACHCAR, J. A. An alternative reparametrization for the weighted lindley distribution. **Pesquisa Operacional**, SciELO Brasil, v. 36, n. 2, p. 345–353, 2016. Citations on pages 26 and 41.

MAZUCHELI, J.; COELHO-BARROS, E. A.; LOUZADA, F. On the hypothesis testing for the weighted lindley distribution. **Chilean Journal of Statistics**, v. 7, n. 2, p. 17–27, 2016. Citation on page 22.

MAZUCHELI, J.; FERNANDES, L. B.; OLIVEIRA, R. P. de; MAZUCHELI, M. J. **Package 'LindleyR'**. 2016. Citations on pages 22 and 34.

MAZUCHELI, J.; LOUZADA, F.; GHITANY, M. Comparison of estimation methods for the parameters of the weighted lindley distribution. **Applied Mathematics and Computation**, Elsevier, v. 220, p. 463–471, 2013. Citation on page 22.

MCCULLAGH, P.; NELDER, J. Generalized linear models., 2nd edn.(chapman and hall: London). **Standard book on generalized linear models**, 1989. Citation on page 84.

MCKEAGUE, I. W.; SASIENI, P. D. A partly parametric additive risk model. **Biometrika**, Oxford University Press, v. 81, n. 3, p. 501–514, 1994. Citation on page 23.

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**, Brazilian Statistical Association, v. 29, n. 1, p. 132–144, 2015. Citations on pages 24 and 96.

NAKAGAMI, M. The m-distribution—a general formula of intensity distribution of rapid fading. In: **Statistical methods in radio wave propagation**. [S.l.]: Elsevier, 1960. p. 3–36. Citation on page 22.

NASH, J. C.; VARADHAN, R.; GROTHENDIECK, G.; NASH, M. J. C.; YES, L. **Package 'optimx'**. 2020. Citations on pages 66 and 68.

NELSON, W. Theory and applications of hazard plotting for censored failure data. **Technometrics**, Taylor & Francis, v. 14, n. 4, p. 945–966, 1972. Citation on page 31.

NIELSEN, G. G.; GILL, R. D.; ANDERSEN, P. K.; SØRENSEN, T. I. A counting process approach to maximum likelihood estimation in frailty models. **Scandinavian journal of Statistics**, JSTOR, p. 25–43, 1992. Citation on page 39.

NIELSEN, H. B.; MORTENSEN, S. B. **ucminf: General-Purpose Unconstrained Non-Linear Optimization**. [S.l.], 2016. R package version 1.1-4. Available: <https://CRAN.R-project.org/package=ucminf>. Citations on pages 66 and 85.

NOCEDAL, J.; WRIGHT, S. Numerical optimization springer-verlag. **New York**, 1999. Citations on pages 66, 85, and 98.

OLCAY, A. H. Mean residual life function for certain types of non-monotonic ageing. **Communications in statistics. Stochastic models**, Taylor & Francis, v. 11, n. 1, p. 219–225, 1995. Citation on page 48.

ORTEGA, E. M.; CORDEIRO, G. M.; CAMPELO, A. K.; KATTAN, M. W.; CANCHO, V. G. A power series beta weibull regression model for predicting breast carcinoma. **Statistics in medicine**, Wiley Online Library, v. 34, n. 8, p. 1366–1388, 2015. Citations on pages 24, 25, 37, 38, and 99.

PARNER, E. *et al.* Inference in semiparametric frailty models. **ACTA JUTLANDICA**, AARHUS UNIVERSITY PRESS, v. 73, p. 320–321, 1998. Citation on page 39.

PATIL, G. *et al.* The weighted distribution: A survey of their applications. 1977. Citation on page 32.

PATIL, G. P.; RAO, C. R. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. **Biometrics**, JSTOR, p. 179–189, 1978. Citation on page 32.

PHINYO, P.; PATUMANOND, J.; PONGUDOM, S. Time-dependent treatment effects of metronomic chemotherapy in unfit aml patients: a secondary analysis of a randomised controlled trial. **BMC Research Notes**, Springer, v. 14, n. 1, p. 1–6, 2021. Citation on page 23.

PICKLES, A.; CROUCHLEY, R. A comparison of frailty models for multivariate survival data. **Statistics in Medicine**, Wiley Online Library, v. 14, n. 13, p. 1447–1461, 1995. Citation on page 40.

PIEGORSCH, W. W. Maximum likelihood estimation for the negative binomial dispersion parameter. **Biometrics**, JSTOR, p. 863–867, 1990. Citation on page 80.

PRENTICE, R. L. Linear rank tests with right censored data. **Biometrika**, Oxford University Press, v. 65, n. 1, p. 167–179, 1978. Citation on page 23.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. Available: <https://www.R-project.org/>. Citations on pages 34, 45, 52, 66, 68, 85, 86, and 99.

RAFTERY, A. E.; NEWTON, M. A.; SATAGOPAN, J. M.; KRIVITSKY, P. N. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. bepress, 2006. Citation on page 46.

RAMOS, P.; LOUZADA, F. The generalized weighted lindley distribution: Properties, estimation, and applications. **Cogent Mathematics**, Taylor & Francis, v. 3, n. 1, p. 1256022, 2016. Citation on page 23.

RAMOS, P. L.; ALMEIDA, M. P.; TOMAZELLA, V. L.; LOUZADA, F. Improved bayes estimators and prediction for the wilson-hilferty distribution. **Anais da Academia Brasileira de Ciencias**, SciELO Brasil, v. 91, n. 3, 2019. Citations on pages 22, 55, and 56.

RAMOS, P. L.; LOUZADA, F.; CANCHO, V. G. Maximum likelihood estimation for the weighted lindley distribution parameters under different types of censoring. **Revista Brasileira de Biometria/Biometric Brazilian Journal**, v. 35, n. 1, p. 115–131, 2017. Citation on page 23.

RAO, C. R. On discrete distributions arising out of methods of ascertainment. **Sankhyā: The Indian Journal of Statistics, Series A**, JSTOR, p. 311–324, 1965. Citation on page 32.

RATNANINGSIH, D. J.; SAEFUDDIN, A.; KURNIA, A. Stratified-extended cox with frailty model for non-proportional hazard: A statistical approach to student retention data from universitas terbuka in indonesia. **Thailand Statistician**, v. 19, n. 1, p. 209–228, 2021. Citation on page 23.

REED, W. J. A flexible parametric survival model which allows a bathtub-shaped hazard rate function. **Journal of Applied Statistics**, Taylor & Francis, v. 38, n. 8, p. 1665–1680, 2011. Citation on page 55.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005. Citation on page 91.

RIGBY, R. A.; STASINOPOULOS, M. D.; HELLER, G. Z.; BASTIANI, F. D. **Distributions for modeling location, scale, and shape: Using GAMLSS in R**. [S.l.]: CRC press, 2019. Citation on page 26.

ROBERT, C.; CASELLA, G. **Monte Carlo statistical methods**. [S.l.]: Springer Science & Business Media, 2013. Citation on page 40.

ROCHA, R.; NADARAJAH, S.; TOMAZELLA, V.; LOUZADA, F. Two new defective distributions based on the Marshall-Olkin extension. **Lifetime Data Analysis**, Springer Verlag, New York, v. 22, p. 216–240, 2016. Citation on page 63.

RODRIGUES, J.; CANCHO, V. G.; CASTRO, M. de; LOUZADA-NETO, F. On the unification of long-term survival models. **Statistics & Probability Letters**, Elsevier, v. 79, n. 6, p. 753–759, 2009. Citations on pages 25, 29, 37, 38, 79, and 80.

RODRIGUES, J.; CASTRO, M. de; BALAKRISHNAN, N.; CANCHO, V. G. Destructive weighted poisson cure rate models. **Lifetime data analysis**, Springer, v. 17, n. 3, p. 333–346, 2011. Citation on page 25.

RODRIGUES, J.; CASTRO, M. de; CANCHO, V. G.; BALAKRISHNAN, N. Com–poisson cure rate survival models and an application to a cutaneous melanoma data. **Journal of Statistical Planning and Inference**, Elsevier, v. 139, n. 10, p. 3605–3611, 2009. Citation on page 38.

ROSS, G.; PREECE, D. The negative binomial distribution. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 34, n. 3, p. 323–335, 1985. Citation on page 84.

SAHA, K.; PAUL, S. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. **Biometrics**, Wiley Online Library, v. 61, n. 1, p. 179–185, 2005. Citation on page 80.

SANTOS-NETO, M.; CYSNEIROS, F. J. A.; LEIVA, V.; AHMED, S. E. On new parameterizations of the birnbaum-saunders distribution. **Pakistan Journal of Statistics**, v. 28, n. 1, 2012. Citation on page 54.

SANTOS-NETO, M.; CYSNEIROS, F. J. A.; LEIVA, V.; BARROS, M. Reparameterized birnbaum-saunders regression models with varying precision. **Electronic Journal of Statistics**, The Institute of Mathematical Statistics and the Bernoulli Society, v. 10, n. 2, p. 2825–2855, 2016. Citation on page 26.

SASIENI, P. D. Proportional excess hazards. **Biometrika**, Oxford University Press, v. 83, n. 1, p. 127–141, 1996. Citation on page 23.

SCHEMPER, M. Cox analysis of survival data with non-proportional hazard functions. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 41, n. 4, p. 455–465, 1992. Citations on pages 23 and 35.

SCHOENFELD, D. Partial residuals for the proportional hazards regression model. **Biometrika**, Oxford University Press, v. 69, n. 1, p. 239–241, 1982. Citation on page 35.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Citation on page 54.

SHANKER, R.; SHUKLA, K. K.; LEONIDA, T. A. Weighted quasi lindley distribution with properties and applications. **International Journal of Statistics and Applications**, Scientific & Academic Publishing Co., v. 9, n. 1, p. 8–20, 2019. Citation on page 23.

SMITH, R. M.; BAIN, L. J. An exponential power life-testing distribution. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 4, n. 5, p. 469–481, 1975. Citation on page 22.

STEWART, J. **Single variable calculus: Early transcendentals**. [S.l.]: Cengage Learning, 2015. Citation on page 31.

STRUTHERS, C. A.; KALBFLEISCH, J. D. Misspecified proportional hazard models. **Biometrika**, Oxford University Press, v. 73, n. 2, p. 363–369, 1986. Citation on page 25.

SUGIURA, N. Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 7, n. 1, p. 13–26, 1978. Citation on page 54.

THOMAS, D. C. Use of auxiliary information in fitting nonproportional hazards models. **Modern statistical methods in chronic disease epidemiology**, Wiley New York, v. 197210, 1986. Citation on page 23.

TIBSHIRANI, R. J.; CIAMPI, A. A family of proportional-and additive-hazards models for survival data. **Biometrics**, JSTOR, p. 141–147, 1983. Citation on page 23.

TSODIKOV, A.; IBRAHIM, J.; YAKOVLEV, A. Estimating cure rates from survival data: an alternative to two-component mixture models. **Journal of the American Statistical Association**, Taylor & Francis, v. 98, n. 464, p. 1063–1078, 2003. Citation on page 25.

TSODIKOV, A. D.; YAKOVLEV, A. Y.; ASSELAIN, B. **Stochastic models of tumor latency and their biostatistical applications**. [S.l.]: World Scientific, 1996. Citation on page 24.

VASQUEZ, J. K.; RODRIGUES, J.; BALAKRISHNAN, N. A useful variance decomposition for destructive waring regression cure model with an application to hiv data. **Communications in Statistics-Theory and Methods**, Taylor & Francis, p. 1–12, 2020. Citations on pages 25 and 38.

VAUPEL, J. W.; MANTON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. **Demography**, Springer, v. 16, n. 3, p. 439–454, 1979. Citations on pages 25 and 26.

VAUPEL, J. W.; YASHIN, A. I. The deviant dynamics of death in heterogeneous populations. RR-83-001, 1983. Citation on page 26.

VENABLES, W. N.; RIPLEY, B. D. **Modern applied statistics with S-PLUS**. [S.l.]: Springer Science & Business Media, 2013. Citation on page 66.

VILCA, F.; SANTANA, L.; LEIVA, V.; BALAKRISHNAN, N. Estimation of extreme percentiles in birnbaum–saunders distributions. **Computational statistics & data analysis**, Elsevier, v. 55, n. 4, p. 1665–1678, 2011. Citations on pages 73 and 75.

WIENKE, A. **Frailty models in survival analysis**. [S.l.]: CRC press, 2010. Citations on pages 21, 23, 25, 26, 29, 30, 34, 38, 39, 40, 62, 63, and 73.

WILSON, E. B.; HILFERTY, M. M. The distribution of chi-square. **proceedings of the National Academy of Sciences of the United States of America**, National Academy of Sciences, v. 17, n. 12, p. 684, 1931. Citation on page 22.

YAKOVLEV, A. Y.; TSODIKOV, A. D.; BASS, L. A stochastic model of hormesis. **Mathematical Biosciences**, Elsevier, v. 116, n. 2, p. 197–219, 1993. Citations on pages 24, 25, 38, and 83.

YIQI, B.; RUSSO, C. M.; CANCHO, V. G.; LOUZADA, F. Influence diagnostics for the weibull-negative-binomial regression model with cure rate under latent failure causes. **Journal of Applied Statistics**, Taylor & Francis, v. 43, n. 6, p. 1027–1060, 2016. Citation on page 91.

YU, J. Empirical characteristic function estimation and its applications. **Econometric reviews**, Taylor & Francis, v. 23, n. 2, p. 93–123, 2004. Citation on page 46.

ZAKERZADEH, H.; DOLATI, A. Generalized lindley distribution. JOURNAL OF MATHEMATICAL EXTENSION, 2009. Citation on page 21.


ICMC
USP
SÃO CARLOS