UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA – UFSCar-USP

Juliana Scudilio Rodrigues

Cure Rate Models:
Alternatives Methods to Estimate the Cure Rate

Thesis submitted to the Department of Statistics – DEs/UFSCar and to the Institute of Mathematics and Computer Sciences – ICMC-USP in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Doctor in Statistics.

Advisor: Prof. Dra. Vera Lúcia Damasceno Tomazella

São Carlos
September 2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA – UFSCar-USP

Juliana Scudilio Rodrigues

Modelos de Fração de Cura: Métodos Alternativos para Estimar a Proporção de Curados

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

Orientador: Prof. Dra. Vera Lúcia Damasceno Tomazella

São Carlos
Setembro de 2020

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 da candidata Juliana Scudilio Rodrigues, realizada em 16/07/2020.

Comissão Julgadora:

Profa. Dra. Vera Lucia Damasceno Tomazella (UFSCar)

Prof. Dr. Eder Angelo Milani (UFG)

Profa. Dra. Giovana Oliveira Silva (UFBA)

Profa. Dra. Gleici da Silva Castro Perdoná (USP)

Prof. Dr. Vinicius Fernando Calsavara (CIPE)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

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

ACKNOWLEDGEMENTS

À Instituição pelo ambiente criativo e amigável que proporciona. A Universidade Federal de São Carlos (UFSCAR) e a Universidade de São Paulo (USP-IME) em especial ao Departamento de Estatística de ambas as universidades, pela oportunidade de fazer o doutorado em Estatística.

A professora Dra. Vera Tomazella pela orientação, por acreditar em mim e pelo empenho dedicado na elaboração desta tese.

Ao professor Dr. Gustavo Henrique de Araújo Pereira e Dr.Vinicius Casalvara pela parceria de pesquisa, por todo ensinamento, pelas conversas e conselhos ao longo desses anos. A banca examinadora pela sugestões oportunas e correções feitas durante a defesa.

Agradeço a todos os professores do departamento de estatística da USP e UFSCAR por me proporcionar o conhecimento não apenas racional, mas a manifestação do caráter e afetividade da educação no processo de formação profissional, por tanto que se dedicaram a mim, não somente por terem me ensinado, mas por terem me feito aprender. A palavra mestre, nunca fará justiça aos professores dedicados aos quais sem nominar terão os meus eternos agradecimentos.

A minha família, pelo amor, incentivo e apoio incondicional. Em especial ao Ricardo Rocha por todo companheirismo e apoio.

Aos meus amigos de doutorado que fizeram esses anos serem mais fáceis. A todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

"Do not worry about your difficulties in Mathematics.
I can assure you mine are still greater."
(Albert Einstein)

RESUMO

SCUDILIO, J. . **Modelos de Fração de Cura: Métodos Alternativos para Estimar a Proporção de Curados**. 2020. 94 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, 2020.

Modelos com fração de cura formaram um importante campo de pesquisa na área de análise de sobrevivência e tem atraído a atenção dos pesquisadores. Na busca de novos modelos de fração de cura, esse trabalho tem como principal objetivo propor métodos alternativos para modelar a proporção de curados. Neste contexto apresentamos dois métodos alternativos a metodologia existente na literatura. O primeiro método tem enfoque em modelos defeituosos, os quais têm a vantagem de modelar a proporção de cura sem adicionar parâmetros extras no modelo, em contraste com a maioria dos modelos da literatura. Este método propõem modelos defeituosos induzidos por fragilidade gama, nessa abordagem mostramos que podemos induzir novas distribuições defeituosas ao usar o termo de fragilidade gama. Modelos com termos de fragilidade incorporam uma heterogeneidade não observada entre os indivíduos e a incorporação dessa heterogeneidade não observada traz vantagens para a estimação dos modelos. O segundo método proposto utiliza famílias de distribuições para calcular a fração de cura. Nesta abordagem incluímos um parâmetro na família de distribuição Beta-G e uma nova família de modelos de fração de cura mais flexível para modelagem de dados com fração de cura é considerada.

Palavras-chave: Análise de sobrevivência, Família de distribuições, Modelos defeituosos, Modelos de fragilidade, Modelos de fração de cura, Modelos de longa duração.

ABSTRACT

SCUDILIO, J. . **Cure Rate Models: Alternatives Methods to Estimate the Cure Rate**. 2020. 94 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, 2020.

Cure rate models in survival data studies has formed an important field in the area and has attracted the attention of researchers. In the search for new models of cure rate, the objective of this work is to propose alternative methods to model the cure rate. For this two methods are presented. The first use the methodology of the defective models and the last method use the concept the distributions family. Then, in the first method propose the defective models induced by a frailty term. Defective models have the advantage of modeling the proportion of cured without adding any extra parameters in the model, in contrast to the most models from the literature. Models with a frailty term incorporate an unobserved heterogeneity among individuals and this incorporation brings advantages for the estimated model, because it incorporates the influence of unobserved covariates in a proportional hazard model. It is showed that the new defective distributions are induced when using the gamma frailty term.

The last method proposed in this work, is to use distribution families to calculate the cure rate. For this, a parameter ”*p*” is included in the Beta-G family in order to create a new family of cure rate models, the new family can be more flexible for modeling cure rate than the standard mixture models.

Keywords: Cure rate models, Defective models, Family distributions, Frailty models, Long-term survivors, Survival Analysis.

LIST OF FIGURES

Figure 1 – Example: the cumulative density function of a defective distribution. 34

Figure 2 – The density probability, the survival and the hazard functions of defective distribution. 35

Figure 3 – Density probability function, survival function and hazard function of the inverse Gaussian distribution. 36

Figure 4 – Mean squared errors, bias, coverage probabilities and coverage length of $(\hat{a}, \hat{\beta}_0, \hat{\beta}_1, \hat{p}_0, \hat{p}_1)$ versus n for the defective Gompertz model. The red, green and blue lines (squares, circles and triangles) represent scenarios I, II and III, respectively. 39

Figure 5 – Mean squared errors, bias, coverage probabilities and coverage length of $(\hat{a}, \hat{\beta}_0, \hat{\beta}_1, \hat{p}_0, \hat{p}_1)$ versus n for the defective inverse Gaussian model. The red, green and blue lines (squares, circles and triangles) represents the scenario where, respectively. 40

Figure 6 – Estimated cumulative hazard curves for colon, divorce, melanoma and breast cancer data respectively. 42

Figure 7 – KM curve and estimated curve of the Gompertz (left) and inverse Gaussian (right) models for the colon cancer data. 43

Figure 8 – Fitted survival curves of the Gompertz (left graphic) and inverse Gaussian models (right graphic). the divorce data set 44

Figure 9 – Kaplan-Meier curves for the variables TIL, N and T. 45

Figure 10 – Estimated survival curves of the Gompertz model (left graphic) and the inverse Gaussian model (right graphic) in breast cancer data. 46

Figure 11 – Estimated survival curves of the Gompertz model (left graphic) and the inverse Gaussian model (right graphic) in melanoma data. 47

Figure 12 – Graphics of the cure fraction of the Gompertz and inverse Gaussian varying the frailty parameter θ. 51

Figure 13 – Mean squared errors, bias, coverage probabilities and covarage length of $(\hat{a}, \hat{\theta}, \hat{\beta}_0, \hat{\beta}_1, \hat{p}_0, \hat{p}_1)$ versus n for the defective gamma-Gompertz model. The red, green and blue lines (squares, circles and triangles) represents the scenario where $\theta = 0.75, 1, 2$, respectively. 58

Figure 14 – Mean squared errors, bias, coverage probabilities and covarage length of $(\hat{a}, \hat{b}, \hat{\theta}, \hat{\beta}_0, \hat{\beta}_1, \hat{p}_0, \hat{p}_1)$ versus n for the the defective gamma-inverse Gaussian model. The red, green and blue lines (squares, circles and triangles) represents the scenario where $\theta = 0.75, 1, 2$, respectively. 60
Figure 15 – Estimated survival curves of the gamma-Gompertz model and frailty mixture gamma-Gompertz for the colon data set. 62
Figure 16 – Estimated survival curves of the gamma-Gompertz model and frailty mixture gamma-Gompertz model with the covariate ethnicity in divorce data. 63
Figure 17 – Estimated survival curves of the gamma-inverse Gaussian model and frailty mixture gamma-inverse Gaussian model with the covariate N. 64
Figure 18 – Estimated survival curves of the gamma-inverse Gaussian model with three covariates. In the left, the survival curves when TIL=0 and N=1. In the right, the survival curves when TIL=1 and N=1. 64
Figure 19 – Estimated survival curves of the gamma-inverse Gaussian model and frailty mixture gamma-inverse Gaussian for the melanoma data set. 66
Figure 20 – Curves for the probability density (first line), survival function (second line) and hazard (third line) functions of the Beta-pExp distribution for several parameter choices. 73
Figure 21 – Curves for the density (first line), survival (second line) and hazard (third line) functions of the Beta-pLindley distribution for several parameter choices. 74
Figure 22 – Simulation of the Beta-pExp distribution. The red, green and blue (square, circle and triangle) represent scenarios one, two and three, respectively. . . . 78
Figure 23 – AIC and BIC mean difference between the mixture Beta-Exp and Beta-pExp distributions. The left, mid and right panels represents the first, second and third scenarios, respectively. 79
Figure 24 – Cure rate estimates for each group obtained from both models. The left, mid and right panels represents the first, second and third scenarios, respectively. 79
Figure 25 – Kaplan–Meier survival estimates curves (black line) stratified by treatment (0: control group; 1: treatment group) and survival function estimates according to different models. 81
Figure 26 – Kaplan–Meier survival estimates curves (black line) stratified by sex covariate and survival function estimates according to different models. 83

LIST OF TABLES

Table 1 – Maximum likelihood estimates, standard error, 95% confidence intervals (CIs), AIC and BIC criteria according to Gompertz and inverse-Gaussian models for the colon data. 42

Table 2 – Maximum likelihood estimates, standard error, 95% confidence intervals (CIs), AIC and BIC criteria according to Gompertz and inverse-Gaussian models for the divorce data. 44

Table 3 – Descriptive of each variable in the breast cancer data set. The $\{0,1\}$ values in the columns stands for the censored and failure times, respectively. The $\{0,1\}$ value in the lines represents the covariate categories. 45

Table 4 – Maximum likelihood estimates, standard error, 95% confidence intervals (CIs), AIC and BIC criteria according to Gompertz and inverse-Gaussian models for the breast cancer data. 46

Table 5 – Maximum likelihood estimates, standard error, 95% confidence intervals (CIs), AIC and BIC criteria according to Gompertz and inverse-Gaussian models for the melanoma data. 47

Table 6 – Maximum likelihood estimates, standard error (SE), 95% confidence intervals (CIs), AIC and BIC criteria according to gamma-Gompertz and frailty mixture gamma-Gompertz for the colon cancer data. 61

Table 7 – Maximum likelihood estimates, standard error (SE), 95% confidence intervals (CIs), AIC and BIC criteria according to gamma-Gompertz and frailty mixture gamma-Gompertz for the divorce data. 63

Table 8 – Maximum likelihood estimates, standard error (SE), 95% confidence intervals (CIs), AIC and BIC criteria according to gamma-inverse Gaussian and frailty mixture gamma-inverse Gaussian for the breast cancer data. with one covariate (N). 63

Table 9 – Maximum likelihood estimates of the gamma-inverse Gaussian model with three covariates. 65

Table 10 – Estimated cure rates for each configuration of covariates. 65

Table 11 – Maximum likelihood estimates, standard error (SE), 95% confidence intervals (CIs), AIC and BIC criteria according to gamma-inverse Gaussian and frailty mixture gamma-inverse Gaussian for the melanoma cancer data. 66

Table 12 – Maximum likelihood estimates. standard error (SE). 95% confidence interval and AIC values obtained by Beta-pExp and Mixture Beta-Exp models fits for the melanoma data set. 81
Table 13 – Maximum likelihood estimates. standard error (SE). 95% confidence interval and AIC. BIC. CAIC values obtained by Beta-pExp and Mixture Beta-Exp models fits for the oncocentro data set. 82
Table 14 – Breast cancer data set collected at the hospital A.C.Camargo Cancer Center. 94

CONTENTS

1 INTRODUCTION 19
1.1 Objective and Overview 21

2 BACKGROUND 23
2.1 Survival analysis 23
2.2 Kaplan Meier estimator 25
2.3 Cure rate model 25
2.4 Frailty model 27
2.5 Maximum Likelihood Estimation 28
2.6 Delta Method 30
2.7 Selection Models 30
2.8 Artificial Data Generation Algorithm 30
2.9 Conclusion 32

3 DEFECTIVE MODELS 33
3.1 Defective Distributions 33
3.1.1 Gompertz Defective Distribution 34
3.1.2 Inverse Gaussian Defective Distribution 35
3.2 Defective regression model 36
3.2.1 Defective Gompertz regression model 36
3.2.2 Defective inverse Gaussian regression models 36
3.3 Inference 37
3.4 Simulation studies 38
3.5 Application 41
3.5.1 Colon data 42
3.5.2 Divorce data 43
3.5.3 Breast cancer data 44
3.5.4 Melanoma data 46
3.6 Conclusions 48

4 DEFECTIVE MODELS INDUCED BY A FRAILTY TERM 49
4.1 Defective models induced by frailty term 49
4.1.1 Defective gamma-Gompertz model 52
4.1.2 Defective gamma-inverse Gaussian model 53

4.1.3 Defective regression models induced by a frailty term 54
4.1.3.1 Defective gamma-Gompertz regression model 55
4.1.3.2 Defective gamma-inverse Gaussian regression model 55
4.1.4 Inference 56
4.2 Simulation studies 57
4.3 Application 61
4.3.1 Colon cancer 61
4.3.2 Divorce data 62
4.3.3 Breast cancer 62
4.3.4 Melanoma 65
4.4 Conclusions 66

5 BETA-PG FAMILY 69
5.1 The Beta-G Family of Distributions 69
5.2 The Beta-pG Family of Distributions 70
5.3 The Beta-pG Family of Regression Models 71
5.4 Especial cases 72
5.4.1 The Beta-pWeibull distribution 72
5.4.2 The Beta-pLindley distribution 72
5.4.3 The Beta-pGamma distribution 75
5.5 Inference 75
5.6 Simulation Studies 77
5.7 Applications 80
5.7.1 Melanoma data 80
5.7.2 Oncocentro cancer data 82
5.8 Conclusions 83

6 FINAL REMARKS 85
6.1 Conclusions 85
6.2 Future Works 86
6.3 Acknowledgments 86

BIBLIOGRAPHY 87

APPENDIX A BREAST CANCER DATA SET 93

CHAPTER 1

INTRODUCTION

In survival analysis studies, a cure rate is common. The cure rate is the proportion of the observed individuals who, for some reason, are not susceptible to the event of interest. In different areas, such as in the social area, data sets with this cure rate are observed, like in studies of the occurrence of divorces, or even in studies of the time until the first child birth, or between children, in Medicine, in the study of cancer recurrence and others. An alternative approach that incorporates the existence of cured elements in model is then necessary.

The standard mixture model proposed by Boag (1949) and Berkson and Gage (1952) is the most commonly parametric model used to model the cure rate. In this model, the survival function is given by

$$S(t) = p + (1 - p)S_0(t), \tag{1.1}$$

being $p \in (0,1)$ and $S_0(\cdot)$ is a proper survival function.

Thus, $S(t)$ converges to p when the time increases. The most common choices for $S_0(\cdot)$ are the survival functions of the Weibull, log-logistic and log-normal. Others parametric approach can be found in Rodrigues *et al.* (2009) who proposed an unified long-term theory that generalizes, among others, the standard mixture model and the promotion time model of Chen, Ibrahim and Sinha (1999).

Defective distributions are another way to model a cure fraction, a concept formalized by Balka, Desmond and McNicholas (2009). In defective models, the cure proportion p is not directly estimated, as in the standard mixture model, it was used a distribution that naturally becomes a cure rate model, when changing the usual domain of its parameters. The integral of the density function in defective model is not 1, but a value in the range $(0,1)$. Then, the survival function converges to a value of $p \in (0,1)$, leading to a cure fraction. And the distribution is no longer proper.

The cure rate of the population in defective models is obtained by calculating the limit of the survival function using the estimated parameters. In contrast to the usual models of cure rate, defective models do not need any assumption about the existence of the cure rate previously to the modeling, since these models have a natural structure of cure rate model. In the literature, there are two distributions for this purpose: the inverse Gaussian and Gompertz. The Gompertz model is used to fit breast cancer data (HAYBITTLE, 1959) and a modified version is used to fit pediatric cancer data (CANTOR; SHUSTER, 1992). The Gompertz model with covariates was proposed by Gieser *et al.* (1998). The cure rate model with inverse Gaussian distribution was proposed in the papers of Balka, Desmond and McNicholas (2009), Balka, Desmond and McNicholas (2011) and Whitmore (1979), .

A Bayesian approach of the defective Gompertz model and its comparison with the maximum likelihood estimation were proposed by Rocha, Tomazella and Louzada (2014) and Santos, Achcar and Martinez (2017). Martinez and Achcar (2017) proposed the generalized Gompertz models and Borges (2017) used the EM algorithm in the defective generalized Gompertz regression models. An extension of the Gompertz and inverse Gaussian models using the Marshall-Olkin family of distributions was proposed by Rocha *et al.* (2015b), the Marshall-Olkin Gompertz and Marshall-Olkin Gaussian inverse, respectively. It is shown that these models can also take defective versions. The Kumaraswamy family is used to extend the Gompertz and inverse Gaussian distribution, with applications in cancer data studies (ROCHA *et al.*, 2015a). A new way to generate defective distributions was proposed based on a Marshall-Olkin family new property (ROCHA *et al.*, 2017). The authors used the extended Weibull class of distributions combined under the Marshall-Olkin family to generate ten new defective distributions, as examples. Many other authors have been working with defective distributions, for example, Martinez and Achcar (2018) proposed the defective Dagun distribution (DDD). Calsavara *et al.* (2019c) proposed a defective regression model for survival data modeling with a proportion of early failures or zero-adjusted. Also, a cure rate defective model for interval-censored event-time data was proposed by Calsavara *et al.* (2019b).

In survival analysis, its common for two elements with the same characteristics to present responses at different times. For example, in medical studies, two patients with the same characteristics do not have the same medical response at the same time, i.e., there are biological variables that are not being measured between these individuals that justify this fact. This heterogeneity exists for many reasons, some of which are said to be unobserved variability, such as environmental factors, genetic factors or information not collected (LAURIE *et al.*, 1989). Hougaard (1991) showed that it would be advantageous to consider two sources of heterogeneity: the observed (covariates) and unobserved.

The frailty model proposed by Vaupel, Manton and Stallard (1979) considers this variability in the lifetime. In this model a random effect is included, i.e., is included a random variable representing the information that cannot or has not been observed. One way to incorporate this frailty is to introduce into the hazard function, in order to control the heterogeneity of variables.

The frailty term can be included in a multiplicative or an additive form and, in this work, it was considered the multiplicative frailty models, which were also considered by many papers.

A review of multiplicative frailty models from the classical point of view was presented by Hougaard (1995) and Andersen *et al.* (1993). And a review of these models from a Bayesian perspective was presented by Sinha and Dey (1997). Vaupel, Manton and Stallard (1979) used the frailty term in uni-variate data, Clayton (1978) and Oakes (1982) used it in multivariate models. The extend of frailty models considering a cure fraction was proposed by Hougaard, Myglegaard and Borch-Johnsen (1994), Jr and Halloran (1996), Price and Manatunga (2001), Aalen and Gjessing (2001), Yau and Ng (2001), Peng, Taylor and Yu (2007), Yu and Peng (2008), Calsavara *et al.* (2017) and Calsavara *et al.* (2019a).

One of the objectives of this thesis is to present a new and more flexible defective distribution. This new defective distribution is induced by a frailty term with gamma distribution. It will be provided the implementation of a frailty term in a defective model and its estimation process.

Another approach to estimate the cure rate is to use the distribution families. These families add extra parameters in a baseline distribution in order to increase its modeling capability. Some family of distributions mentioned are: the Beta-G family distributions (EUGENE; LEE; FAMOYE, 2002), the exponential-G family distributions (GUPTA; GUPTA; GUPTA, 1998), the gamma-G family distributions (ZOGRAFOS; BALAKRISHNAN, 2009),the Kumaraswamy-G family distributions (CORDEIRO; CASTRO, 2011), the generalized Beta-G family distributions (ALEXANDER *et al.*, 2012), the Beta extended-G family distributions (CORDEIRO; ORTEGA; SILVA, 2012), the Beta exponential-G family distributions and the Weibull-G family distributions (ALZAATREH; LEE; FAMOYE, 2013), exponential generalized-G family distributions due to Cordeiro, Ortega and Cunha (2013), among others. Then, another objective of this work is to propose a new way to include the parameter p in family distribution, in order to create a new family of cure rate models, which is here shown that adds more flexibility to modeling than the standard mixture approach.

Thus, in the search of the new models to estimate the proportion of cured, the main goal of this thesis is to propose two alternative methods to estimate the cure rate.

1.1 Objective and Overview

The cure rate model is an important area in survival analysis and many methodologies have been proposed in this model. The defective model is a cure rate model that has the advantage of not assuming the presence of immune elements.

The main objective of this work is to propose methods to estimate the proportion of cured using different methodologies, especially the methods that use the defective distribution.

The specific objectives are listed below:

- To propose a defective distribution induced by a gamma frailty term.
- To propose a family Beta-pG using a family distribution to include the parameter "p" in order to create a new family of cure rate models.

The overview of this work is written in independent chapters with each chapter presenting a new research contribution. The chapters are related in the sense that all chapter talk about the cure rate models. In the first method, defective models is used; in the last method, the family distribution is used.

Thereby, thesis is organized as follows: first in Chapter 2, a brief review of the concepts of survival analysis is presented, as well as the cure rate model and the frailty model in the literature. Furthermore, in Chapter 3, is presented the defective distributions and the defective regression models. To finalize this chapter, simulation studies and four applications in real data set are presented.

In Chapter 4, the defective models with the frailty term are formulated and its properties are discussed. Simulation studies are presented to analyze the asymptotic properties of maximum likelihood estimators. To finalize this chapter, three real databases are presented to illustrate the proposed distribution compared to the standard mixture models. This chapter is based on the paper of Scudilio *et al.* (2019).

Chapter 5 proposes the Beta-pG family of distributions, a family of distributions to model cure rates in survival data. The proposed family is defined and some of its details are discussed. Moreover, it is presented an approach to introduce covariates in this model and discuss the estimation procedure by maximum likelihood. Simulation studies are presented to analyze the asymptotic properties of the maximum likelihood estimators and to compare the proposed model with the mixture models. To finalize this chapter, we present two applications in real data sets to illustrate the proposed methodology. Both data sets are related to cancer. Finally, in Chapter 6, the conclusions of this work and some proposals for future works are discussed.

CHAPTER 2

BACKGROUND

The objective of this chapter is to present a background about survival analysis and the principal models of cure rate models. In this brief review, is presented in Section 2.1 concepts of survival analysis. In Section 2.3, the cure rate models are defined and the special cases are presented. Section 2.4 presents the frailty models. In the next sections, the estimation by maximum likelihood for the cure rate models is presented (Section 2.5), as well as the delta method is used to calculate the variance associated the cure rate parameter (Section 2.6), and the measures to selection models (Section 2.7). Finally, in Section 2.8 presented the algorithm to generate the artificial data.

2.1 Survival analysis

In survival analysis studies, the interest is to estimate the time until the event of interest. For example, in the industry, the time until the equipment fails; in the social area, the time until the divorce or the birth of the first child; in medicine, the time can be the lifetime of a patient or the time to cure a disease.

The presence of incomplete observations (or censored observations) is a characteristic in survival analysis data set. Censored observations can consist of loss of follow-up with the patient, lack of knowledge about the onset of the disease, death of the patient from another cause, etc. These are all examples of right censoring; in this work, only right censored was used.

There are three types of right censored observations: censored type I, censored type II and random censored, which follow:

- Type I censoring occurs when the study is followed up until a predetermined time,
- Type II censoring consists of finishing the study by obtaining a predetermined number of censoring,

- Random censoring occurs when the patient withdraws from the study without presenting the event of interest.

In survival analysis each observation is denoted by (t_i, δ_i), being t_i the time until fail or censoring, and δ_i the censoring indicator. If $\delta_i = 1$, failure is observed. Then, if $\delta_i = 0$, the observation is censored.

Then, considering T a non-negative random variable, being $T > 0$, with probability density function $f(t)$, the density function is defined by

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}. \tag{2.1}$$

Then, the probability density function (pdf) is the limit of the probability of an individual failing in the time interval $[t, t + \delta t]$.

The cumulative density function is given by

$$F(t) = P(T \leq t) = \int_0^t f(u)du. \tag{2.2}$$

The survival function is the probability of an individual's survival to the time t, then

$$S(t) = P(T > t) = \int_t^\infty f(u)du = 1 - F(t). \tag{2.3}$$

The function $S(t)$ presents this property:

i. $S(t)$ is a non increasing function,

ii. $S(0) = 1$;

iii. $\lim_{t \to \infty} S(t) = 0$,

If property [iii] is satisfied, the it is said that $S(t)$ is a proper survival function.

The hazard function is another function that is important in survival analysis. The hazard function is the instantaneous rate at which events occur for individuals who are surviving at time t, in other words, what is the chance of an individual to fail at the time $t + \Delta t$, with $\Delta \to 0$. The hazard function is given by

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

To finalize, the cumulative hazard function is another interesting function defined as follows

$$H(t) = \int_0^t h(u)du. \tag{2.4}$$

There are useful relations between these functions, as

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} = -\frac{d}{dt}\log[S(t)], \\ H(t) &= -log[S(t)], \\ S(t) &= \exp[-H(t)]. \end{aligned}$$

2.2 Kaplan Meier estimator

The Kaplan-Meier (KM) estimator or product-limit Kaplan is the most important nonparametric estimator in survival analysis, and it is used to estimate the survival function. (KAPLAN; MEIER, 1958). The KM estimator is defined as

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right). \tag{2.5}$$

where

- $t_{(1)} < t_{(2)} < \ldots < t_k$, being k the distinct and ordered fail times;
- d_j is the number of fail in $t_{(j)}$, $j = 1, \ldots, k$;
- n_j is the number of individuals at risk in time $t_{(j)}$, which means, individuals who have not failed or were not censored.

Kaplan and Meier (1958) showed that the KM estimator is the maximum likelihood estimate for $S(t)$. For that, the KM curve is widely to verify the lack of fit of the proposed parametric survival model.

2.3 Cure rate model

The long-term models or cure rate models are an important part of survival analysis and have been widely used in different areas. The unified class of the cure rate model proposed by Rodrigues *et al.* (2009) is here briefly described.

The Unified models are a class of cured fraction model that generalizes the mixture model and the promotion model, among others. The latent variable is proposed to represent the number of the competitive causes of an event of interest. The cure rate models incorporate the heterogeneity of two sub populations: the susceptible population and the immune population (cured) to the event of interest. We define this class as follows.

Let N be a random variable that represents the number of competitive causes for a particular event of interest with probability

$$p_n = P[N = n],$$

where $n = 0, 1, 2, \ldots$. In this case, N is a latent variable. Let $Z_v | N = n$, $v = 1, \ldots, n$, independent and non-negative random variable with a cumulative density function given by $F(t) = 1 - S(t)$, that does not depend of N.

The random variables Z_v represent the time until the occurrence of a particular event of interest for the risk cause v. Let $T = min\{Z_1, Z_2, \ldots, Z_n\}$ be the time until the event of interest occurs, in which $P(Z_0 = \infty) = 1$ leads to a proportion p_0 of the non-susceptible observations of the event of interest. The random variable T is a censure or an observable random variable and Z_v are latent variables. The survival function of the random variable T is defined as $S(t) = P(T > t)$.

So, let $\{a_n\}$ be a sequence of real numbers in $s\,\varepsilon\,[0,1]$. Then

$$A(s) = a_0 + a_1 s + a_2 s^2 + \ldots,$$

converges and $A(s)$ is defined as a generating function of the sequence $\{a_n\}$. (See proof in Feller (1968)).

According to Rodrigues *et al.* (2009), given a proper survival function, $S_0(t)$, the survival function of the random variable T is given by

$$S_{pop}(t) = A[S_0(t)] = \sum_{n=0}^{\infty} p_n \left[S_0(t)\right]^n. \tag{2.6}$$

The cure rate, p, is given by $\lim_{t \to \infty} S_{pop}(t) = P(N = 0) = p$. Observe that the survival function $S_{pop}(t)$, defined in (2.6), is not a proper function. The probability density function and the hazard function are given by

$$\begin{aligned} f_{pop}(t) &= f_0(t)\frac{d}{ds}A[S_0(t)], \\ h_{pop}(t) &= \frac{f_{pop}(t)}{S_{pop}(t)} = \frac{f_0(t)}{S_1(t)}\frac{d}{dt}A[S_0(t)]. \end{aligned}$$

The likelihood function is given by

$$L(\boldsymbol{\gamma}; D) = \prod_{i=1}^{n} [f_{pop}]^{\delta_i} [S_{pop}(t)]^{(1-\delta_i)} = \prod_{i=1}^{n} \left[f_0(t)\frac{d}{ds}A[S_0(t)]\right]^{\delta_i} \left[\sum_{n=0}^{\infty} p_n \left[S_0(t)\right]^n\right]^{(1-\delta_i)}.$$

The following distributions can be used as generating functions, $A(s)$: Bernoulli, Binomial, Poisson, Negative Binomial, Geometric, among others. Whereas N follows a Bernoulli distribution, then $A[s] = p + (1-p)s$ and then is obtained the standard mixture model.

The standard mixture model is given by

$$S_2(t) = p + (1-p)S_0(t), \tag{2.7}$$

where p represents the proportion of cured population and $S_0(t)$ is the baseline survival function.

The promotion model is given when N follows a Poisson distribution with mean θ, then $A[s] = \exp(-\theta + \theta s)$. Thus, the survival function of the promotion model is given by

$$\begin{aligned} S_3(t) &= P(N=0) + P(Z_1 > t, \ldots, Z_N > t, N \geq 1) \\ &= \exp(-\theta) + \textstyle\sum_{k=1}^{\infty} S_0(t)^k \frac{\theta^k}{k!} \exp(-\theta) \\ &= \exp(-\theta F_0(t)). \end{aligned} \tag{2.8}$$

It's important to emphasize that the survival functions 2.8 is not proper, because $S_3(\infty) = \exp(-\theta) > 0$.

2.4 Frailty model

The frailty model is a model of random effects for time variables. In this model a random effect is included, i.e., is included a random variable representing the information that cannot or has not been observed. One way to incorporate this frailty is to introduce into the hazard function, in order to control the heterogeneity of variables.

Models with frailty terms incorporate an unobserved heterogeneity among individuals, which represents advantages for model estimation.

The multiplicative frailty term model is an extension of Cox model (COX, 1972), which is the most commonly used model to incorporate the term of frailty. In these models, the risk depends on the unobserved and non-negative random variable V, that includes in a multiplicative way the basic risk function. Thus, considering the models with a frailty term in a multiplicative way in the risk function, the frailty models are defined as:

Let $F_0(t)$ be the cumulative function distribution, $S_0(t)$ a proper or a not proper survival function, and $h_0(t)$ the respective hazard function and considering a nonnegative unobservable random variable V that denotes the frailty term. Then, the conditional hazard function with a frailty term is defined by

$$h(t|V) = V\,h_0(t). \tag{2.9}$$

In equation (2.9), V is the frailty term, when V increases, the risk of failure also increases.

A problem in frailty models is the choice of the distribution of the random effect. The frailty distribution most often applied is the gamma distribution (CLAYTON, 1978), (VAUPEL; MANTON; STALLARD, 1979), (MISSOV, 2010), (MISSOV, 2013). However, other choices can be considered, such as the positive stable distribution (HOUGAARD, 1986b), a three-parameter distribution (PVF) (HOUGAARD, 1986a), the compound Poisson distribution (AALEN, 1992), the log-normal and inverse Gaussian distributions (TOMAZELLA, 2003), among others.

If the distribution of the random effect must have expectation 1, then the model will be identifiable (ELBERS; RIDDER, 1982). In this work, it is assumed that the random variable V follows a gamma distribution with shape parameter k and inverse scale parameter λ ($V \sim \Gamma(k,\lambda)$), with $\mathbb{E}(V) = k/\lambda$ and $\text{Var}(V) = k/\lambda^2$, (WIENKE, 2003). Here, it is considered that $\mathbb{E}(V) = k/\lambda = 1$ and $\text{Var}(V) = k/\lambda^2 = \theta$, where $k = \lambda = \theta^{-1}$.

Given the frailty model presented in equation (2.9), the conditional cumulative hazard function and the conditional survival function are define as

$$\begin{aligned} H(t|V) &= VH_0(t), \\ S(t|V) &= e^{-vH_0(t)} = [S_0(t)]^v. \end{aligned}$$

And the unconditional survival function is defined as

$$S(t) = \mathbb{E}[S(t|V)] = \int_0^\infty S(t|v)g(v)dv = \int_0^\infty e^{-VH_0(t)}g(v)dv, \tag{2.10}$$

where $g(v)$ is the density function of frailty term.

One way to solve the equation 2.10 consists of using the Laplace transform.

Definition 2.4.1. Laplace transform: The Laplace transform of the function $f : [0,\infty] \to \mathbb{R}$

$$\mathscr{L}_f(s) = \int_0^\infty e^{-st} f(t)dt, \tag{2.11}$$

$\forall s \geq 0$ in that the integral 2.11 converges.

Observe that equation (2.10) is similar to equation (2.11). Using the Laplace transform, the unconditional survival function is obtained, being defined by

$$S(t) = \mathbb{E}[S(t|V)] = \mathbb{E}\left[e^{-H(t|V)}\right] = \mathbb{E}\left[e^{-VH_0(t)}\right] = \mathscr{L}\left[H_0(t)\right].$$

It is important to mention that the Laplace transform is very important in the frailty model.

Using the derivatives of Laplace transform, the unconditional density function $f(t)$ and risk function $h(t)$ are obtained. They are

$$f(t) = -h_0(t)\mathscr{L}'(H_0(t)), \tag{2.12}$$

$$h(t) = -h_0(t)\frac{\mathscr{L}'(H_0(t))}{\mathscr{L}(H_0(t))}. \tag{2.13}$$

2.5 Maximum Likelihood Estimation

The most common method to estimate the parameters in cure rate models is the maximum likelihood estimator. It has good property in large samples (asymptotic result) and it is possible to incorporate the censorship observed of the data set.

Assuming that the data are independent, identically distributed and have a density and survival function denoted by $f(.,\boldsymbol{\theta})$ and $S(.,\boldsymbol{\theta})$, respectively, where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k)^T$ is a vector of parameters. Consider a data set $\boldsymbol{D} = (\boldsymbol{t},\boldsymbol{\delta})$, where $\boldsymbol{t} = (t_1,\ldots,t_n)^T$ are the observed failure times and $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_n)^T$ are the censored failure times, being $\delta_i = 1$ when the failure is observed and 0 otherwise.

So, the likelihood function of $\boldsymbol{\theta}$ is defined by

$$L(\boldsymbol{\theta},\boldsymbol{D}) \propto \prod_{i=1}^{n} f(t_i,\boldsymbol{\theta})^{\delta_i} S(t_i,\boldsymbol{\theta})^{1-\delta_i}.$$

And the corresponding log-likelihood function is

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta},\boldsymbol{D}) = const + \sum_{i=1}^{n} \delta_i log(f(t_i,\boldsymbol{\theta})) + \sum_{i=1}^{n}(1-\delta_i)S(t_i,\boldsymbol{\theta}).$$

The value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$, or equivalent $l(\boldsymbol{\theta})$, is the maximum likelihood estimator. The estimator is found by solving these equations systems

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{2.14}$$

$U(\boldsymbol{\theta})$ is score function.

Normally, it is necessary to use the numeric methods to solve equation 2.14. In this thesis, R software optim packages was used, and the 'BFGS' method was used to maximize the likelihood function, for details see the R Core Team (2013) package.

Confidence intervals for the parameters are based on the asymptotic normality properties of the maximum likelihood estimators. If $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimators of $\boldsymbol{\theta}$ then the distribution of $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ can be approximated by a multivariate normal distribution k with mean equal to zero and co-variance matrix $I^{-1}(\hat{\theta})$, where $I(\boldsymbol{\theta})$ is the observed information matrix, defined by

$$I(\boldsymbol{\theta}) = -\begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_2 \theta_1} & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_2^2} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_k \theta_1} & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_k \theta_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_k^2} \end{pmatrix}.$$

So, an approximate $100(1-\alpha)\%$ percent confidence interval for θ_i is

$$IC(\boldsymbol{\theta}) = (\hat{\theta}_i - z_{\alpha/2}\sqrt{I^{ii}}, \hat{\theta}_i + z_{\alpha/2}\sqrt{I^{ii}})$$

where I^{ii} denote the ith diagonal element of the inverse of I evaluated at $\hat{\boldsymbol{\theta}}$ and z_α is the $100(1-\alpha)$ percentile of the standard normal distribution. In the models proposed in this thesis, the cure rate

p is calculated as a function of other parameters estimated. Then, to calculate the variance of p, is necessary to use the delta method with a first order Taylor's approximation (OEHLERT, 1992). In the next section, this method will be preseted.

2.6 Delta Method

The delta method is an appropriate way to estimate the variance of parameters of non-linear functions of random variables. This method uses the first step of Taylor approximation to expands the function of a random variable and then take the variance.

Theorem 1. *Consider a sequence of random variables $X_1, X_2, \ldots, X_n$ independent and identically distributed (univariate) with finite variance σ^2. By Limits Theorem has*

$$\sqrt{n}(X_n - g(\theta)) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2),$$

where θ and σ^2 are finite valued constants and $\xrightarrow{D}$ denotes the convergence in distribution. Then

$$\sqrt{n}(g(X_n)\theta) \xrightarrow{D} N(0, \sigma^2).$$

Proof. See the Papanicolaou (2009) □

2.7 Selection Models

To check fit quality of theses models, some measures: AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) and CAIC (Consistent Akaike Information Criterion). They are defined as

$$\begin{aligned} AIC &= 2k - 2l(\boldsymbol{\theta}), \\ BIC &= klog(n) - 2l(\boldsymbol{\theta}), \\ CAIC &= k[log(n) + 1] - 2l(\boldsymbol{\theta}), \end{aligned}$$

where k is the number of parameters in the model, n is the sample size and $l(\boldsymbol{\theta})$ is log-likelihood value in the estimated parameters.

These measures provide a way of model selection, when the model that has the better fit is the one with the lowest AIC or BIC or CAIC. For more details see (AKAIKE, 1974), (SCHWARZ *et al.*, 1978)

2.8 Artificial Data Generation Algorithm

In this section, the data generation used to verify the properties of the maximum likelihood estimator for the distributions presented in this thesis is described. These properties were evaluated here by simulation.

The algorithm described by Rocha *et al.* (2017) was used in all chapters to simulate the data.

Then suppose that the occurrence time of an event-of-interest has cumulative distribution $F(t)$. The main objective is to simulate a random sample with size n containing real times, censored times, p_0 and p_1 the value of the cure fraction for each group. The algorithm is described bellow

Algorithm 1 – Generator of artificial data

```
procedure DATA($\boldsymbol{t}; \boldsymbol{\delta}$)
    Determine the parameter values $\boldsymbol{\theta}$, as well the value of the cure fraction for each group $p_0$ and $p_1$;
    Generate the covariate $X \sim Bin(n, p = 0.5)$;
    for $t \leftarrow 1$ to $n$ do
        if $X = 0$ then, generate $M_i \sim$ Bernoulli $(1 - p_0)$
            if $M_i = 0$ then $t_i' = \infty$.
            end if
            if $M_i = 1$ then $t_i'$ is the root of $F(t) = u$, where $u \sim$ uniform$(0, 1 - p_0)$;
            end if
        end if
        if $X = 1$ then, generate $M_i \sim$ Bernoulli $(1 - p_1)$;
            if $M_i = 0$ then $t_i' = \infty$.
            end if
            if $M_i = 1$ then $t_i'$ is the root of $F(t) = u$, where $u \sim$ uniform$(0, 1 - p_1)$;
            end if
        end if
    end for
    for $t \leftarrow 1$ to $length(t_i')$ do
        if $t_i' \neq \infty$ then generate $u_i' \sim Unif(0, max(t_i'))$
        end if
        Calculate $t_i = \min(t_i', u_i')$.
        if $t_i < u_i'$ then $\delta_i = 1$
        else $\delta_i = 0$.
        end if
    end for
    return $(\boldsymbol{t}, \boldsymbol{\delta})$
end procedure
```

In simulation, the value of 1000 simulation per sample size was choosed. In each sample, it is calculated the bias, mean square error, coverage probability and coverage lengths for each parameter, defined as following

$$\begin{aligned} Var(\hat{\theta}) &= \frac{1}{S}\sum_{i=1}^{S}(\hat{\theta}_i - \theta)^2, \\ Bias(\hat{\theta}) &= \hat{\theta} - \theta, \\ MSE(\hat{\theta}) &= Var(\hat{\theta}) + Bias^2(\hat{\theta}) \end{aligned}$$

where $\hat{\theta}$ is the estimator of θ_i for $i = 1, 2, \ldots, S$.

The coverage probability represents the observed percentage of times that the interval contains the true value of the parameters. And the coverage length is the difference between the upper and lower confidence bounds.

2.9 Conclusion

In this chapter, the background about the cure rate model and frailty model were presented. In the next chapter, the defective distribution and the defective models will be presented.

CHAPTER 3

DEFECTIVE MODELS

Defective cure rate models, have the advantage of not assuming the presence of immune elements in the data. These models present a defective distribution as a baseline function and estimate the proportion of cured without adding extra parameters in a model.

Cure rate models via defective distribution is a recent methodology and it is not frequently used in statistics. There are few defective distributions in the literature in this chapter, the basic defective distributions found and the literature are presented: Gompertz distributions and the inverse Gaussian (Section 3.1).

Additionally, an option to include covariates in defective models (Section 3.2) and their estimation process (Section 3.3) are presented. Finally, a simulation studies (Section 3.4) and four real data sets are presented to illustrate the proposed models (Section 3.5).

3.1 Defective Distributions

Definition 3.1.1. Defective distribution: A distribution is called defective if the integral of the density function results in a value $p \in (0,1)$, when the domain of its parameters is changed.

In defective distribution, the cumulative function does not approach 1, but to a value $p \in (0,1)$, and then the survival function approaches $(1-p)$.

The defective model is a model with a defective distribution. These models have the advantage of being unnecessary to assume the existence of a cure fraction. Once you have a defective model, it will lead to a cure fraction when the estimation procedure presents a value out of the usual range of parameters.

The proportion of the immune population is obtained by calculating the limit of the survival function using the estimated parameters. A disadvantage is that these models may lose some flexibility by having fewer parameters.

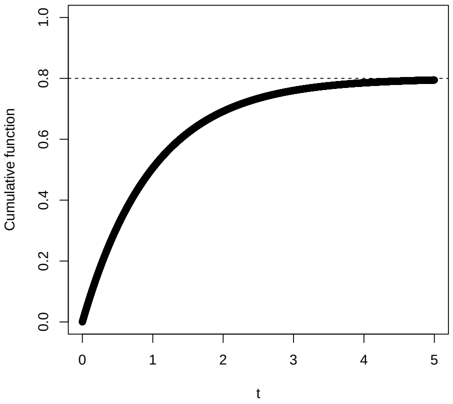

Figure 1 – Example: the cumulative density function of a defective distribution.

Furthermore, since the cure rate depends on other parameters, its interval estimation, it is not obtained directly, and needs to be approximated using other techniques, such as the delta method.

The Gompertz and inverse Gaussian distributions are the only two known common distributions used in defective models. Both distributions have two positive parameters. When the shape parameter assumes negative values, the distribution becomes defective. The parameters that have changed their domain are called defective parameters. In the next section, both distributions are defined.

3.1.1 Gompertz Defective Distribution

The Gompertz distribution is often used to model survival data in various knowledge areas (GIESER *et al.*, 1998). The probability density function for the Gompertz distribution is given by

$$g_0(t) = be^{at}e^{-\frac{b}{a}(e^{at}-1)},$$

where $a > 0$, $b > 0$ and $t > 0$. The corresponding survival function is

$$S_0(t) = e^{-\frac{b}{a}(e^{at}-1)},$$

and the hazard function is given by

$$h_0(t) = be^{at}.$$

When the parameter a assume negative values, we have the defective Gompertz distribution. And the proportion of immunity in the population is calculated as the limit of the survival function, when $a < 0$, given by

$$\lim_{t\to\infty} S_0(t) = \lim_{t\to\infty} e^{-\frac{b}{a}(e^{at}-1)} = e^{\frac{b}{a}} = p \in (0,1).$$

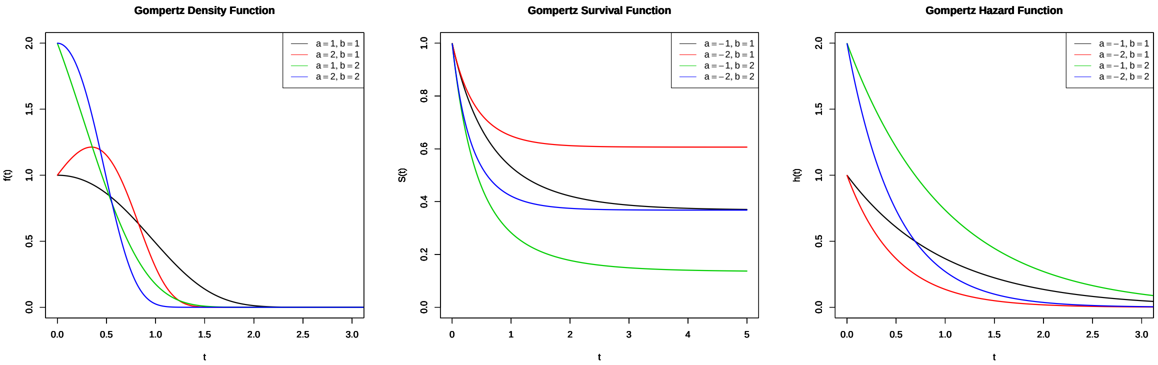

Figure 2 – The density probability, the survival and the hazard functions of defective distribution.

Figure 2 presented the density probability function, survival function and hazard function of the defective Gompertz model using different values for the parameters.

3.1.2 Inverse Gaussian Defective Distribution

Other distribution used to model survival data is the inverse Gaussian distribution. The inverse Gaussian distribution has probability density function given by

$$g_0(t) = \frac{1}{\sqrt{2b\pi t^3}} \exp\left\{-\frac{1}{2bt}(1-at)^2\right\},$$

where $a > 0$, $b > 0$ and $t > 0$. The corresponding survival function is given by

$$S_0(t) = 1 - \left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right],$$

where $\Phi(\cdot)$ denotes the cumulative distribution of the standard normal. The hazard function is

$$h_0(t) = \frac{g_0(t)}{S_0(t)} = \frac{\frac{1}{\sqrt{2b\pi t^3}} \exp\left\{-\frac{1}{2bt}(1-at)^2\right\}}{1 - \left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right]}.$$

The defective inverse Gaussian distribution is the inverse Gaussian distribution that allows negative values of parameter a. The cure rate is calculated by

$$\lim_{t\to\infty} S_0(t) = \lim_{t\to\infty} 1 - \left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right] = (1-e^{2a/b}) = p \in (0,1).$$

Figure 3 shows the density probability, and survival and hazard functions of the defective inverse Gaussian distribution for different values of the parameters.

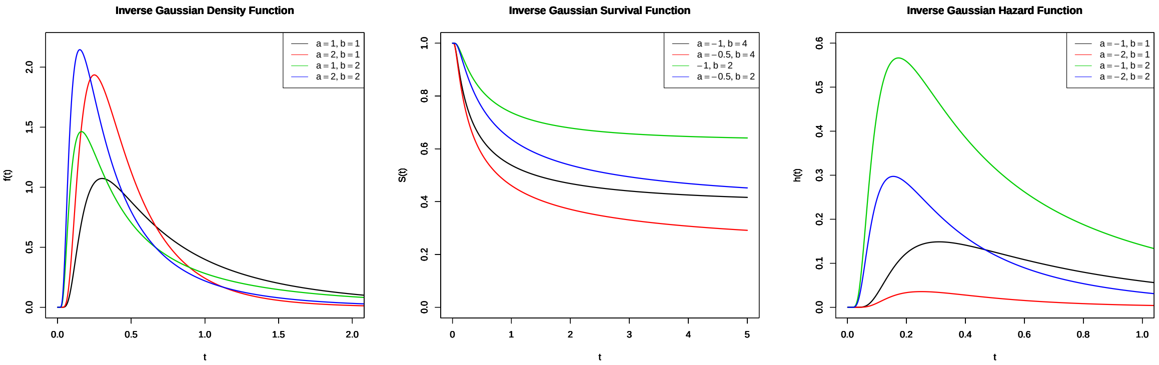

Figure 3 – Density probability function, survival function and hazard function of the inverse Gaussian distribution.

3.2 Defective regression model

The interest relation between the variables is common, i.e., being interested in analyzing the effect that one or more explanatory variables (covariates) causes in the failure time. In this section, the defective Gompertz and inverse Gaussian regression models are presented.

3.2.1 Defective Gompertz regression model

Consider $\boldsymbol{x}^\top = (1, x_1, ..., x_p)$ the covariate information vector and $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, ...\beta_p)$ a vector of regression coefficients. Given the Gompertz survival function $S_0(t)$ defined in equation 3.1, the respective Gompertz regression model is

$$S_0(t|\boldsymbol{x}) = e^{-\frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{a}(e^{at}-1)}, \tag{3.1}$$

for $a > 0$ and $t > 0$.

The density and hazard function is given by

$$\begin{aligned} g_0(t|\boldsymbol{x}) &= e^{at+\boldsymbol{x}'\boldsymbol{\beta}} e^{-\frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{a}(e^{at}-1)}, \\ h_0(t|\boldsymbol{x}) &= e^{at+\boldsymbol{x}'\boldsymbol{\beta}}. \end{aligned}$$

The model 3.1 was proposed in Parreira *et al.* (2007). The authors called it the proportional hazard model with time-depend. In this work, $a > 0$ and the authors were not commented about when $a < 0$. When $a < 0$, we have the defective regression model with cure rate p estimated by

$$p = e^{\frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{a}} \in (0, 1).$$

3.2.2 Defective inverse Gaussian regression models

Consider $\boldsymbol{x}^\top = (1, x_1, ..., x_p)$ the covariate information vector and $\boldsymbol{\beta} = (\beta_0, \beta_1, ...\beta_p)$ a vector of regression coefficients. Given a survival function $S_0(t)$ of a inverse Gaussian distribution, defined in equation 3.1, the survival function of inverse Gaussian regression model is defined as

$$S_0(t|\boldsymbol{x}) = 1 - \left[\Phi\left(\frac{-1+at}{\sqrt{e^{\boldsymbol{x}'\boldsymbol{\beta}}t}}\right) + e^{2a/e^{\boldsymbol{x}'\boldsymbol{\beta}}}\Phi\left(\frac{-1-at}{\sqrt{e^{\boldsymbol{x}'\boldsymbol{\beta}}t}}\right)\right], \tag{3.2}$$

where $\Phi(\cdot)$ is a cumulative distribution of the standard normal.

The density and hazard function is defined as

$$g_0(t|\boldsymbol{x}) = \frac{1}{\sqrt{2e^{\boldsymbol{x}'\boldsymbol{\beta}}\pi t^3}}\exp\left\{-\frac{1}{2e^{\boldsymbol{x}'\boldsymbol{\beta}}t}(1-at)^2\right\},$$

$$h_0(t|\boldsymbol{x}) = \frac{g_0(t)}{S_0(t)} = \frac{\frac{1}{\sqrt{2e^{\boldsymbol{x}'\boldsymbol{\beta}}\pi t^3}}\exp\left\{-\frac{1}{2e^{\boldsymbol{x}'\boldsymbol{\beta}}t}(1-at)^2\right\}}{1-\left[\Phi\left(\frac{-1+at}{\sqrt{e^{\boldsymbol{x}'\boldsymbol{\beta}}t}}\right)+e^{2a/e^{\boldsymbol{x}'\boldsymbol{\beta}}}\Phi\left(\frac{-1-at}{\sqrt{e^{\boldsymbol{x}'\boldsymbol{\beta}}t}}\right)\right]}.$$

When $a < 0$ the models defined in equation 3.2 is defective and our cure rate is

$$p = 1 - e^{2ae^{\boldsymbol{x}'\boldsymbol{\beta}}}. \tag{3.3}$$

3.3 Inference

In this section, it's discussed the estimation procedures of the defective Gompertz and defective inverse Gaussian model. Consider n the sample size and $T_i = \min(W_i, C_i)$ the observed time, where W_i is time failure and C_i the censoring time. Consider δ_i the failure indicator, $\delta_i = 1$ if $T_i = W_i$ and $\delta_i = 0$ otherwise, for $i = 1, \ldots, n$.

The data is represented as $D = (t, \boldsymbol{\delta}, \boldsymbol{X})$, $t = (t_1, \ldots, t_n)^\top$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$ and $\boldsymbol{X}$ is a covariate matrix $n \times p$. Suppose the data is independently and identically distributed and comes from a distribution with density and survival function defined by $f(\cdot, \boldsymbol{\theta})$ and $S(\cdot, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = \left(a, \boldsymbol{\beta}^\top\right)^\top$ is a parameter vector.

The likelihood function of $\boldsymbol{\theta}$ can be written as

$$L(\boldsymbol{\theta}, D) \propto \prod_{i=1}^{n} f(t_i, \boldsymbol{\theta})^{\delta_i} S(t_i, \boldsymbol{\theta})^{(1-\delta_i)}.$$

The corresponding log-likelihood function is

$$\log L(\boldsymbol{\theta}, D) \;=\; const + \sum_{i=1}^{n} \delta_i \log f(t_i, \boldsymbol{\theta}) + \sum_{i=1}^{n} (1-\delta_i) \log S(t_i, \boldsymbol{\theta}).$$

The log-likelihood for the Gompertz defective model to $\boldsymbol{\theta}$ is

$$\begin{aligned}\log L(\boldsymbol{\theta}, D) &= const + \sum_{i=1}^{n} \delta_i \log f(t_i, \boldsymbol{\theta}) + \sum_{i=1}^{n} (1-\delta_i) \log S(t_i, \boldsymbol{\theta}) \\ &= const + \sum_{i=1}^{n} \delta_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + a\sum_{i=1}^{n} \delta_i t_i - \sum_{i=1}^{n} \left\{\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}(e^{at_i}-1)}{a}\right\}.\end{aligned} \tag{3.4}$$

For inverse Gaussian defective model the log-likelihood is given by

$$\begin{aligned}\log L(\boldsymbol{\theta}, D) &= const + \sum_{i=1}^{n} \delta_i \log\left(\frac{1}{\sqrt{2e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \pi t_i^3}} \exp\left\{-\frac{1}{2e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}{}_t} (1-at_i)^2\right\}\right) + \\ & \sum_{i=1}^{n} (1-\delta_i) \log\left(\frac{1}{\sqrt{2e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \pi t^3}} \exp\left\{-\frac{1}{2e^{\boldsymbol{x}_i' \boldsymbol{\beta}}{}_t} (1-at)^2\right\}\right). \qquad (3.5)\end{aligned}$$

The maximum likelihood estimates are obtained by maximizing the log-likelihood function (3.4) and (3.5). The maximization of functions (3.4) and (3.5) are obtained numerically. There are several routines available for numerical maximization. Here, the optim package of software R and use the BFGS method for maximization were used, for details see the package manual optim R Core Team (2013).

Confidence intervals for the parameters are based on the asymptotic normality properties of the maximum likelihood estimators. If $\widehat{\boldsymbol{\theta}}$ denote the maximum likelihood estimators of $\boldsymbol{\theta}$ then the distribution of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ is approximated by a $p+2$-variate normal distribution with mean equal zero and covariance matrix $I^{-1}\left(\widehat{\boldsymbol{\theta}}\right)$, where $I(\boldsymbol{\theta})$ is the observed information matrix.

So, an approximate $100(1-\alpha)$ % percent confidence interval for θ_i is

$$IC(\boldsymbol{\theta}) = \left(\widehat{\theta}_i - z_{\alpha/2}\sqrt{I^{ii}}, \widehat{\theta}_i + z_{\alpha/2}\sqrt{I^{ii}}\right),$$

where I^{ii} denoted the ith diagonal element of the inverse of I evaluated at $\widehat{\boldsymbol{\theta}}$ and z_a is the $100(1-a)$ percentile of the standard normal distribution.

Ahead, simulation studies for the defective regression models are presented.

3.4 Simulation studies

In this section, it was proposed a study with simulated data to check the properties of the maximum likelihood estimator for the defective Gompertz models and the defective inverse Gaussian model. In order to check the results of simulation studies, were estimated the bias, the mean square errors (MSE), the coverage probability of the Wald-type confidence intervals and the coverage lengths for the parameters $\boldsymbol{\theta}^\top = (a, \boldsymbol{\beta})$. The coverage probability represents the observed percentages of the time in which the interval contains the true value of parameters.

To check the maximum likelihood estimates, six simulations using the algorithm previously described in Section 2.8 were proposed to generate data from the the Gompertz and inverse Gaussian distributions. Three scenarios for the defective Gompertz model and three for the defective inverse Gaussian model were presented. The parameters used in the defective Gompertz distribution were $(a, \beta_0, \beta_1) = (-1, 0.5, 0.5)$, the cure fractions p_0 and p_1 are 0.192 and 0.066 in Scenario I (red lines/squares in the Figure 4). For Scenario II (green lines/circles,Figure 4), the β_0 is decreased to -0.5, then the cure fractions p_0 and p_1 are 0.545 and 0.368. And in Scenario

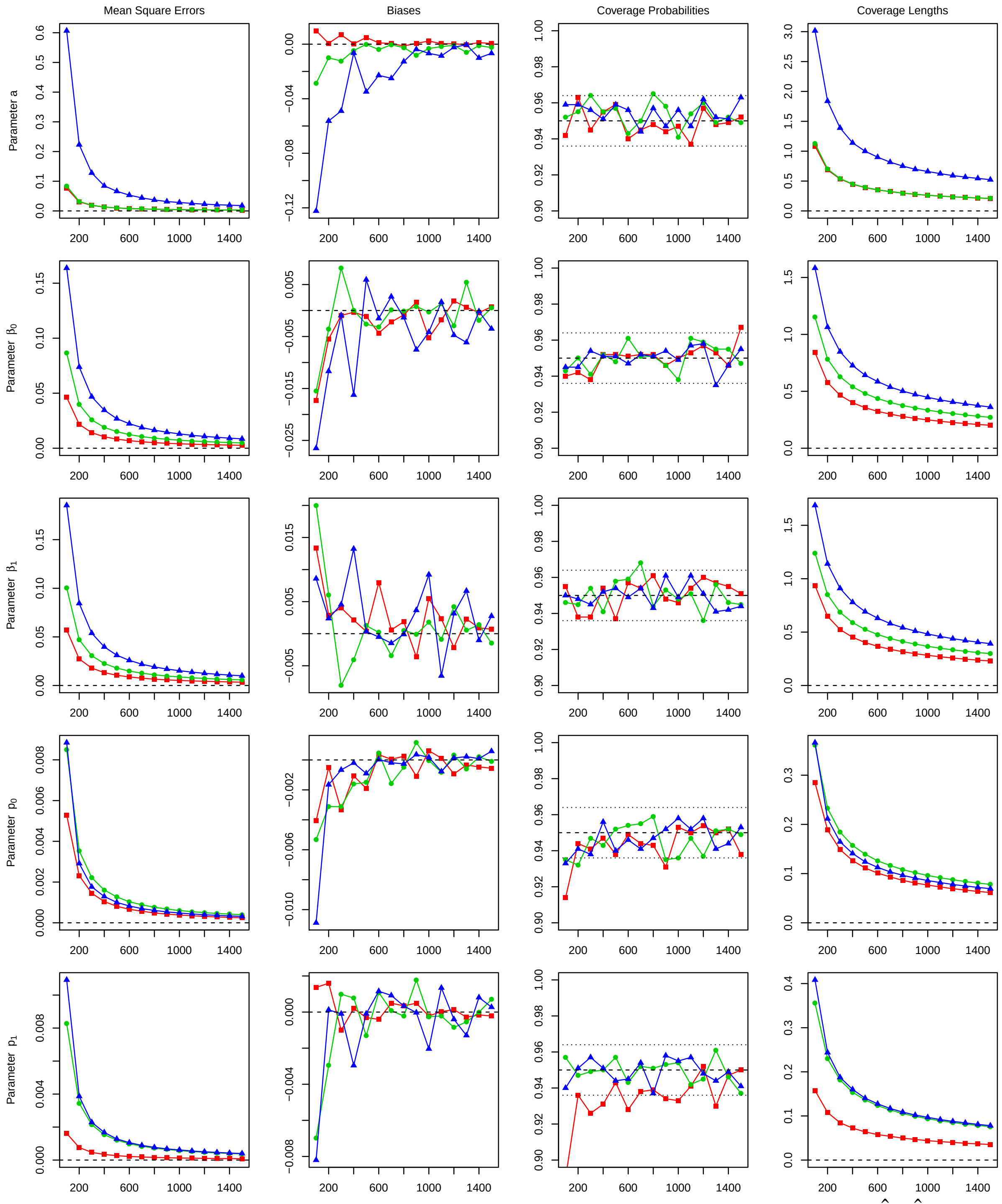

Figure 4 – Mean squared errors, bias, coverage probabilities and coverage length of $(\hat{a}, \hat{\beta}_0, \hat{\beta}_1, \hat{p}_0, \hat{p}_1)$ versus n for the defective Gompertz model. The red, green and blue lines (squares, circles and triangles) represent scenarios I, II and III, respectively.

III (blue lines/triangles, Figure 4), the a decreases to -2.0, so the cure fractions p_0 and p_1 are 0.738 and 0.606.

For the defective inverse Gaussian model, the parameters $(a, \beta_0, \beta_1) = (-1, 0.5,\ 0.5)$ were used, and the cure rate p_0 and p_1 are 0.703 and 0.521 in Scenario I (red lines/squares in the

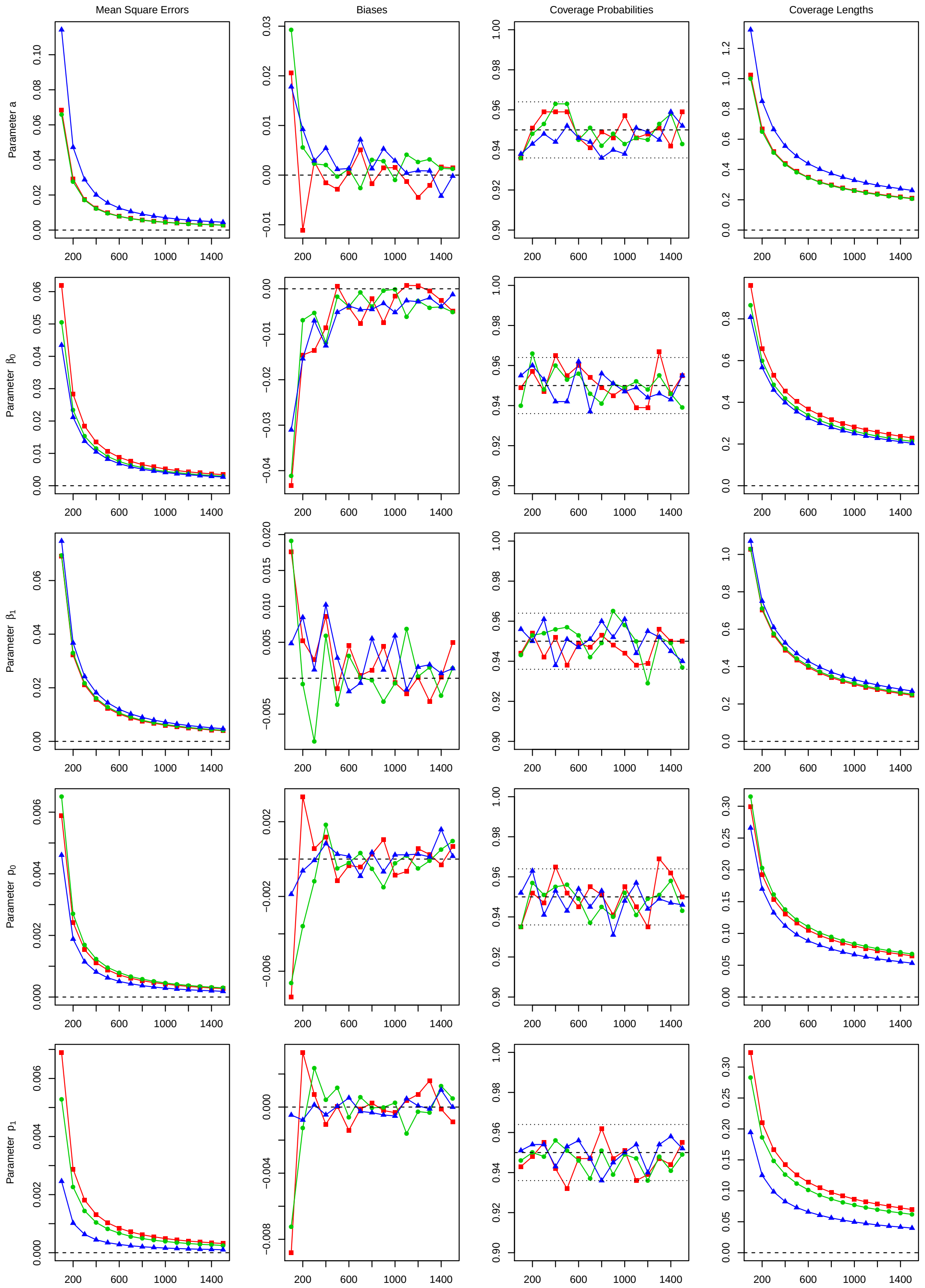

Figure 5 – Mean squared errors, bias, coverage probabilities and coverage length of $(\hat{a}, \hat{\beta_0}, \hat{\beta_1}, \hat{p_0}, \hat{p_1})$ versus n for the defective inverse Gaussian model. The red, green and blue lines (squares, circles and triangles) represents the scenario where, respectively.

Figure 5). In Scenario II (green lines/circles,Figure 5), were considered $(a, \beta_0, \beta_1) = (-1, 1, 0.5)$, with cure rate p_0 and p_1 are 0.521 and 0.360. And for the last scenario (blue lines/triangles, Figure 4), were used $(a, \beta_0, \beta_1) = (-1, 2, 0.5)$, with cure rate $p_0 = 0.237$ and $p_1 = 0.151$.

In each scenario, for each parameter the mean square error, the bias, the coverage probability and range of the confidence interval for different sample sizes were calculated. The sample sizes were considered $n = 100, 200, ..., 1.500$. The delta method was used to estimate the variance of cure fractions. All calculations were developed in R software. Figures 4 and 5 illustrate the results obtained by the defective gamma-Gompertz and defective gamma-inverse Gaussian models, respectively.

From these simulations, it is possible to conclude the following: i) the mean square error decreases and approaches zero as the sample size increases;

ii) the mean square error is smaller for the lowest values of θ;

iii) the mean square error of the cure fractions is generally lower than for the other parameters;

iv) all parameters in all cases converge to their actual values that is, the maximum likelihood estimator of these models is asymptotically not biased;

v) the bias of cure fractions are very small, even in the smallest sample values;

vi) the coverage probabilities are around 0.95 for the parameters a, β_0 , β_1, p_0 and p_1;

vii) the ranges of confidence intervals decrease as the sample size increases;

viii) in general, when the amplitude of the interval is smaller than the other, the smaller the value of θ will be.

In the next section, four applications using the defective regression models showed in this section are presented.

3.5 Application

In this section, four applications are used to exemplify the defective regression models. One of these data set is a social study about divorce. The other data sets ate clinical studies related to cancer occurrence. In all application, used the software R and the function **optim** to estimate the parameter of this models. The same data set is utilized in the chapter 4.

Figure 6 presented the cumulative hazard function for these data sets. In colon data, observe that the cumulative hazard function is constant in initial times, decreases after the time 1000 and in the last times the curve is horizontal, i.e, the hazard is zero. For divorce the data observed an almost constant curve over time, after time 30 the hazard function starts to decline, and in the end presents zero hazard. The curve of the colon data set is similar, but not as smooth, and we have the cumulative hazard function for the breast cancer data set. In the melanoma data set, the hazard starts increasing, then gets constant and start to decrease. Closer to the large event times, the hazard starts to increase again and then goes to zero.

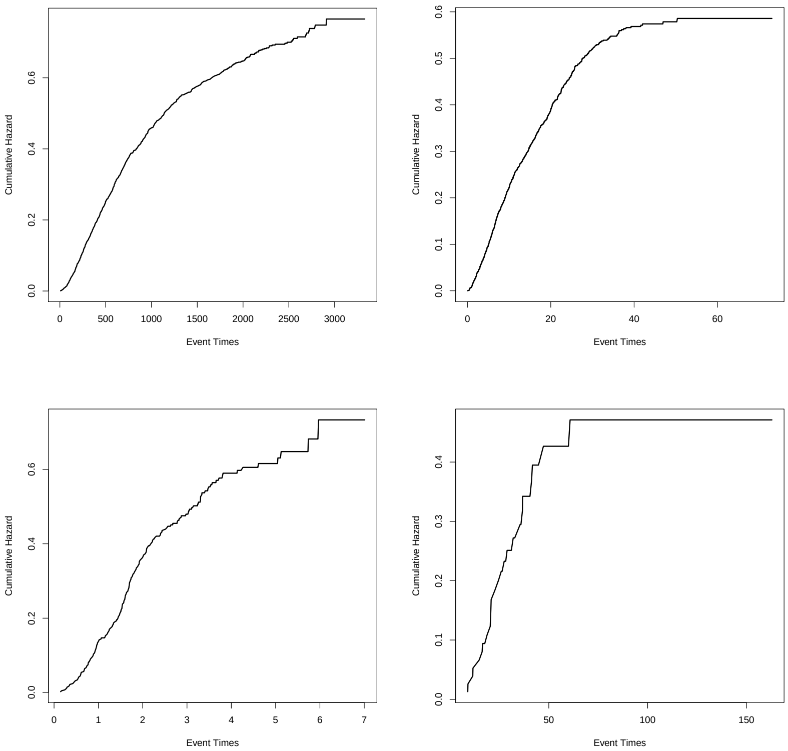

Figure 6 – Estimated cumulative hazard curves for colon, divorce, melanoma and breast cancer data respectively.

3.5.1 Colon data

This data set is one of the first successful studies of chemotherapy for colon cancer. The event of interest is the recurrence or death for the individual under the proposed treatment. The data set has 1858 observations and 938 censures (50.58%) and the covariate used is the existence of adherence to nearby organs. The data set is available in the R, in the package **survival**. The details of this data set can be found in Laurie *et al.* (1989).

Table 1 – Maximum likelihood estimates, standard error, 95% confidence intervals (CIs), AIC and BIC criteria according to Gompertz and inverse-Gaussian models for the colon data.

Distribution	Parameter	Estimate	Std. error	Lower 95% CI.	Upper 95% CI.	AIC	BIC	Log. Veros.
Gompertz	a	-2.3183	0.1772	-2.6656	-1.9709	1507.7300	1524.312	750.8650
	b_0	0.6414	0.0538	0.5359	0.7468			
	b_1	0.3151	0.0868	0.1449	0.4852			
	p_0	0.4408	0.0187	0.4042	0.4774			
	p_1	0.3254	0.0322	0.2623	0.3886			
Inv. Gauss	a	-1.6745	0.1572	-1.9825	-1.3665	1599.102	1615.684	796.5509
	b_0	1.9865	0.0411	1.906	2.0670			
	b_1	0.0536	0.0885	-0.1200	0.2271			
	p_0	0.3683	0.0225	0.3243	0.4123			
	p_1	0.353	0.0294	0.2953	0.4107			

Table 1 summarizes the results for Gompertz and inverse Gaussian regression models in colon data. Observe that both fit, $\hat{a} < 0$, so these models are defective and indicate a cure rate.

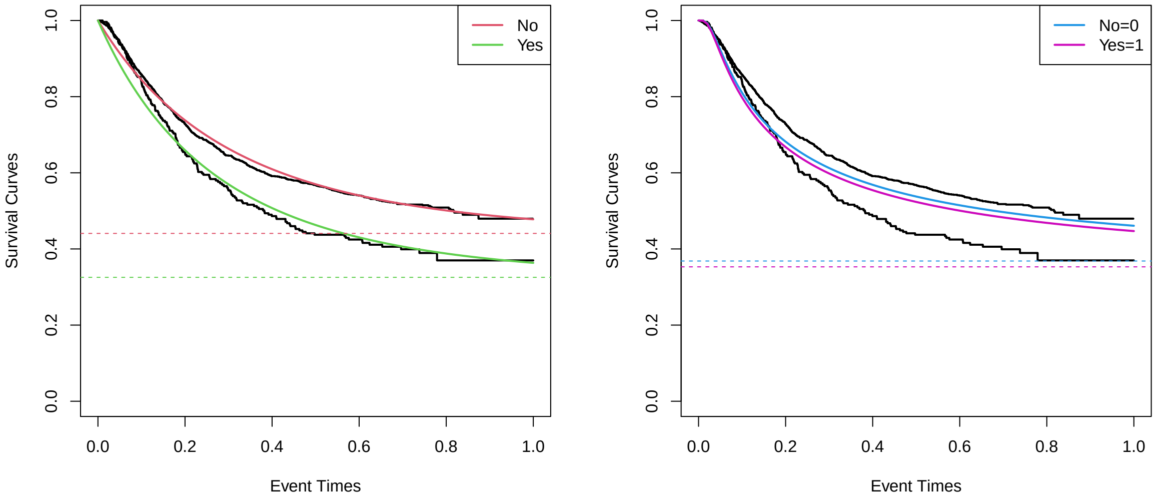

Figure 7 – KM curve and estimated curve of the Gompertz (left) and inverse Gaussian (right) models for the colon cancer data.

The Figure 7 illustrates the survival curve estimated for both models. Observe that the Gompertz regression model (left graphic) fits better than the inverse Gaussian regression model (right graphic). The estimated curve of the Gompertz regression model captures reasonably the Kaplan Meier curve. The Gompertz model presented the smaller AIC and BIC measures.

The proportion of cured were $p_0 = 0.44$, $p_1 = 0.32$ in Gompertz model, i.e, patients who present adherence in the nearby organs have a smaller cure rate than those patients who do not present adherence.

3.5.2 Divorce data

This data set was collected in the United States and observed couples, in which the event of interest is the divorce occurrence. This event may never occur and therefore is high censorship in this data set. The cured elements are those couples who do not get divorced. There are 3371 observations, of which 2339 are censored (69.38%). The maximum time observed was 73.07 years and the mean time observed was 18.41 years. The covariate indicates whether the couple has a different ethnicity. For details on this data, see Lillard and Panis (2000).

Table 2 summarizes the results of the maximum likelihood estimates in Gompertz and inverse Gaussian models for divorce data. Observe that the Gompertz model shows a small advantage when considering AIC, BIC and log-Likelihood measures than the inverse Gaussian model, indicating a better fit. And in both cases $a < 0$, so we have defective models and there is a proportion of elements cured. In both models the cure rate is smaller in couple with different ethnicity, i.e, a couple with different ethnicity has a higher divorce rate than a couple with same ethnicity.

Table 2 – Maximum likelihood estimates, standard error, 95% confidence intervals (CIs), AIC and BIC criteria according to Gompertz and inverse-Gaussian models for the divorce data.

Distribution	Parameter	Estimate	Std. error	Lower 95% CI.	Upper 95% CI.	AIC	BIC	Log. Veros.
Gompertz	a	-2.541	0.230	-2.992	-2.091	1507.719	1526.088	750.860
	b_0	0.588	0.049	0.491	0.685			
	b_1	0.257	0.074	0.111	0.403			
	p_0	0.492	0.023	0.447	0.538			
	p_1	0.400	0.032	0.337	0.463			
Inv.Gauss	a	-3.140	0.204	-3.540	-2.740	1717.153	1735.522	855.577
	b_0	2.138	0.035	2.069	2.206			
	b_1	0.254	0.059	0.138	0.371			
	p_0	0.523	0.017	0.489	0.557			
	p_1	0.437	0.023	0.393	0.481			

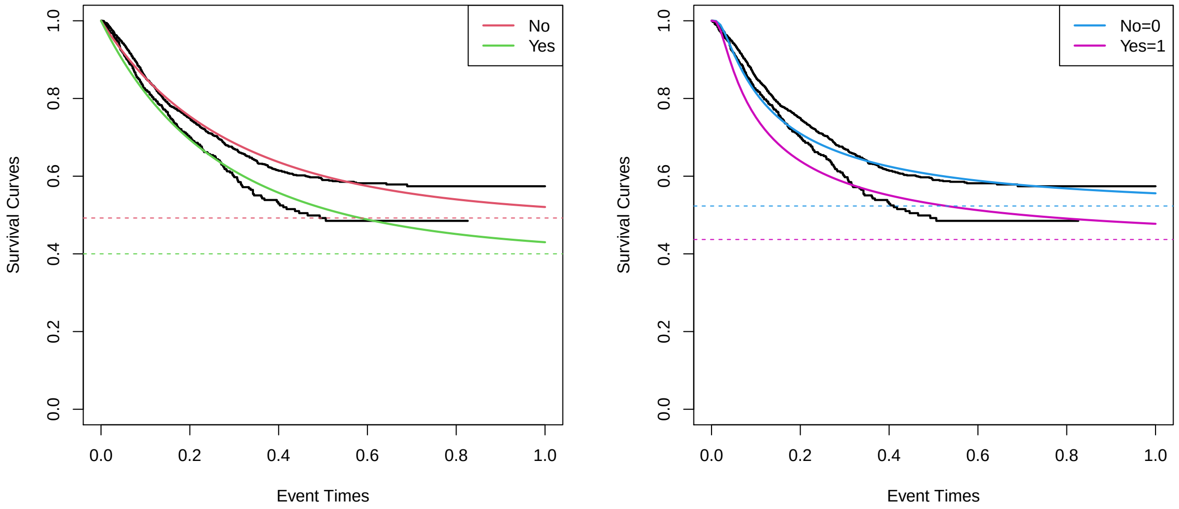

Figure 8 – Fitted survival curves of the Gompertz (left graphic) and inverse Gaussian models (right graphic). the divorce data set

Figure 8 shows the curve fitted survival curves. Note that couples who are not of a different ethnicity have more time until divorce than couples who are of a different ethnicity. That is, couples with ethnicity differences are more likely to divorce than couples with the same ethnicity. The fitted survival curves confirm the poor fit of the proposed models in this data set. Both model fails to capture the behavior of the Kaplan-Meier curves.

3.5.3 Breast cancer data

This data set was collected at the A.C.Camargo Cancer Center, São Paulo, Brazil. It is a study of patients with breast cancer. This data contains information about the failure time or censoring (in months) and three covariates from 78 patients diagnosed with triple-negative breast cancer and treated with neoadjuvant chemotherapy in the period of 2001 to 2013. The event of interest is death by breast cancer. The data set has 53 patients that are censored (67.94%). The three covariates observed was tumor-infiltrating lymphocytes (TIL), the primary tumor site

(T) and the regional lymph node involvement (N). They are all binary variables. The covariate TIL assumes value zero when the patient present TIL$\leq$ 10% and is one when TIL$>$ 10%. The covariate T assume value zero if the tumor characteristic is T0 and is one if the tumor characteristic is T1, T2 or T3. And the covariate N is zero if the nodule characteristic is N0, and is one when the nodule characteristic is N1, N2 or N3. These covariate are defined according to the UICC TNM classifications. This data set has been used for the first time in this paper and it is available in the Appendix.

Table 3 gives descriptive statistics for each one of the covariates. Notice that the covariate TIL contains only 58 observations of the 78 patients in study. These 58 observations, 60.3% of patients present TIL$\leq$ 10%. The covariate T contains 77 information of the 78 patients in the study and in these 77 patients, 37.5% present the tumor characteristic equal T0. The covariate N was observed for all patients, in which 27% present nodule characteristic equal N0. Figure 9 displays the plots of the Kaplan-Meier curve in each variable. In this application, one model only with the covariate N, because is the only one without missing values. For more details about this data see Scudilio *et al.* (2019).

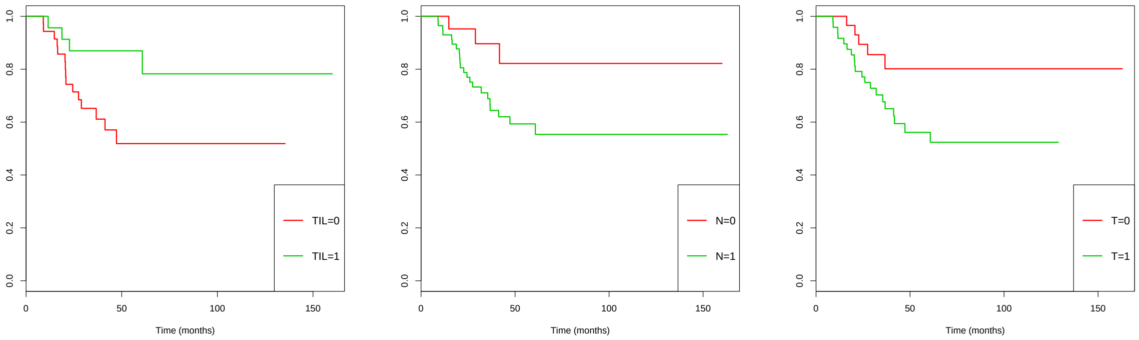

Figure 9 – Kaplan-Meier curves for the variables TIL, N and T.

Table 3 – Descriptive of each variable in the breast cancer data set. The $\{0,1\}$ values in the columns stands for the censored and failure times, respectively. The $\{0,1\}$ value in the lines represents the covariate categories.

TIL[1]	0	1	Total	N[2]	0	1	Total	T[3]	0	1	Total
0	20	15	35	0	18	3	21	0	24	5	29
	51.3%	78.9%	60.3%		34.0%	12.0%	26.9%		46.2%	20.0%	37.7%
1	19	4	23	1	35	22	57	1	28	20	48
	48.7%	21.1%	39.7%		66.0%	88.0%	73.1%		53.8%	80.0%	62.3%
Total	39	19	58	Total	53	25	78	Total	52	25	77
	100.0%	100.0%	100.0%		100.0%	100.0%	100.0%		100.0%	100.0%	100.0%

[1]TIL $=0$ then TIL $\leq 10\%$, TIL $=1$ otherwise. [2]N $=0$ the nodule characteristic is T0; otherwise is N1,N2 or N3. [3]T $=0$ the tumor characteristic is T0; otherwise is T1,T2 or T3

The results of maximum likelihood estimates, standard error, confidence intervals and some measures for Gompertz and inverse Gaussian models are presented in Table 4. The both models fit $a<0$, so we have defective models.

Table 4 – Maximum likelihood estimates, standard error, 95% confidence intervals (CIs), AIC and BIC criteria according to Gompertz and inverse-Gaussian models for the breast cancer data.

Distribution	Parameter	Estimate	Std. error	Lower 95% CI.	Upper 95% CI.	AIC	BIC	Log. Likelihood
Gompertz	a	-1.8166	1.3837	-4.5287	0.8954	45.73517	52.8053	19.86759
	b_0	-0.5615	0.6415	-1.8189	0.6959			
	b_1	1.2015	0.619	-0.0118	2.4148			
	p_0	0.7305	0.1725	0.3924	1.0686			
	p_1	0.352	0.2162	-0.0718	0.7759			
Inv.Gauss	a	-0.5305	0.7528	-2.0061	0.9451	36.29044	43.36057	15.14522
	b_0	0.6278	0.3842	-0.1252	1.3808			
	b_1	0.8076	0.3738	0.0749	1.5403			
	p_0	0.4324	0.3974	-0.3465	1.2113			
	p_1	0.2232	0.2511	-0.2689	0.7152			

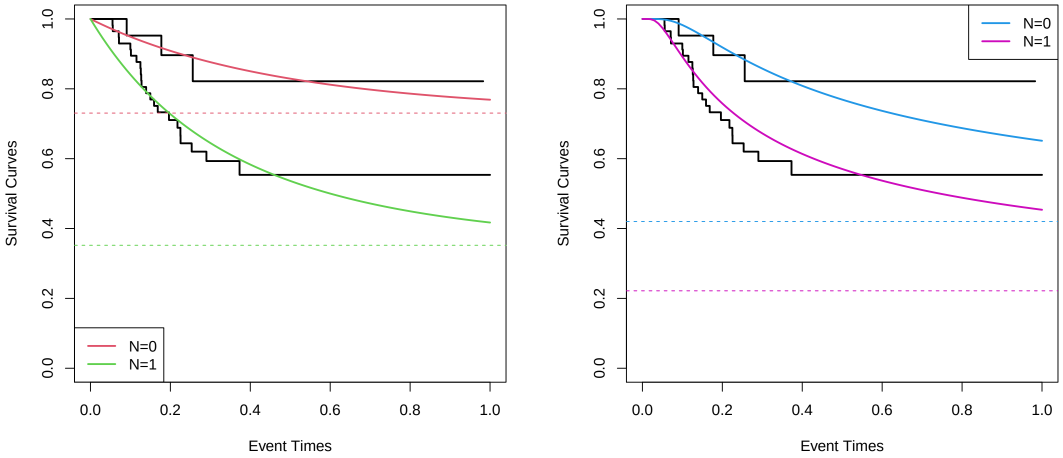

Figure 10 – Estimated survival curves of the Gompertz model (left graphic) and the inverse Gaussian model (right graphic) in breast cancer data.

The inverse Gaussian model presented smaller AIC, BIC and log-likelihood criteria. The proportion of cured estimated in the inverse Gaussian defective model are $p_0 = 0.43$ and $p_1 = 0.22$. In other words, patients classified with N0 nodule have a higher cure rate than those patients with nodule N1, N2 or N3. Figure 10, it can be observed that both models have a poor fit, no models capture the Kaplan-Meier curves well.

3.5.4 Melanoma data

This data set collected in the period 1991-1998 is related to a clinical study in which patients were observed for recurrence after the removal of a malignant melanoma. Melanoma is a type of cancer that develops in melanocytes, responsible for skin pigmentation. It is a potentially serious, malignant tumor that may arise in the skin, mucous membranes, eyes and central nervous system, with a great risk of producing metastases and high mortality rates in the later stages.

There are 417 observed times, of which 232 were censored (55.63%). The covariate is considered to represent the node category, with four categories ($n_1 = 82, n_2 = 87, n_3 = 137, n_4 =$

111). For details of this data, see Ibrahim, Chen and Sinha (2001). Kaplan-Meier estimates (Figure 11) suggest that the cure fraction increases to category nodes.

Table 5 – Maximum likelihood estimates, standard error, 95% confidence intervals (CIs), AIC and BIC criteria according to Gompertz and inverse-Gaussian models for the melanoma data.

Distribution	Parameter	Estimate	Std. error	Lower 95% CI.	Upper 95% CI.	AIC	BIC	Log. Veros.
Gompertz	a	-0.1176	0.0536	-0.2226	-0.0126	1075.748	1095.913	532.8738
	$b0$	-2.1754	0.1943	-2.5562	-1.7946			
	$b12$	0.2664	0.216	-0.157	0.6897			
	$b13$	0.5263	0.2284	0.0786	0.974			
	$b14$	1.0465	0.2146	0.6257	1.4672			
	$p0$	0.3808	0.1444	0.0977	0.6638			
	$p1$	0.2836	0.1372	0.0146	0.5525			
	$p2$	0.1951	0.1247	-0.0494	0.4395			
	$p3$	0.064	0.0687	-0.0706	0.1985			
Inv. Gauss	a	0.0003	0.0294	-0.0574	0.0579	1034.215	1054.381	512.1077
	$b0$	-1.5007	0.1694	-1.8326	-1.1688			
	b12	0.5321	0.1882	0.1632	0.9011			
	$b13$	0.925	0.2132	0.507	1.3429			
	$b14$	1.172	0.2145	0.7516	1.5924			

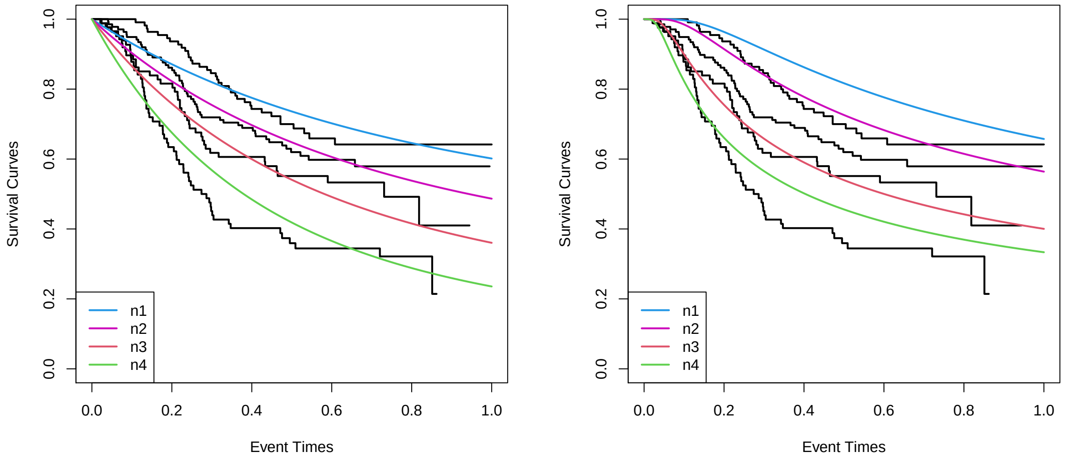

Figure 11 – Estimated survival curves of the Gompertz model (left graphic) and the inverse Gaussian model (right graphic) in melanoma data.

Table 5 shows the maximum likelihood estimates for the Gompertz and inverse-Gaussian models in the melanoma data. In the Gomgpertz model $a < 0$, which indicates defective models, and there is a cure rate. Observe that, in the inverse Gaussian model, the parameter $a > 0$ was very close to zero, indicating a non-cure in this case. Figure 11 illustrates the estimated survival curves. Observe that none of the presented models have a satisfactory fit. In both case, the curve estimates capture poorly the KM curve.

3.6 Conclusions

In this chapter, the concept of defective distributions are introduced and two defective distributions in the literature are presented, the defective Gompertz models and the defective inverse-Gaussian models. It is presented an alternative to add covariates to these defective models. Four applications in real data set were presented to exemplify the defective models presented.

In the applications it becomes clear the need to have more options of defective distributions in the literature. In Rocha *et al.* (2015b) and Rocha *et al.* (2015a) using the distribution families Kumarasawamy and Marshall-Olkin proposed an extension for the defective inverse Gaussian and Gompertz models, thus generating new defective distributions. In Martinez and Achcar (2017) and Martinez and Achcar (2018), the defective generalized Gompertz and the defective Dagun distribution were proposed. In the next chapter, two new defective distributions induced by a frailty term with Gamma distribution will be presented: the defective Gamma-Gompertz and the defective Gaussian Gamma-inverse.

CHAPTER

4

DEFECTIVE MODELS INDUCED BY A FRAILTY TERM

The most adequate way of modeling a frailty term in the presence of immunes elements is using discrete probability distributions. Those distributions allow the value zero for the frailty of an element that have no risk of fail. However, the goal here is to use the gamma distribution for the frailty term to induce a new and more flexible defective distribution. In that case, we won't have a frailty model anymore, but a defective one. This chapter is organized as follows. In Section 4.1 the defective models with the frailty term will be formulated and its properties will be discussed. So two the defective models with the frailty term will be presented, the models defective gamma-Gompertz 4.1.1 and defective gamma-inverse Gaussian 4.1.2. Section 4.1.3, one way to include the covariate in defective models with the frailty term will be presented. Simulation studies are presented to analyze the asymptotic properties of maximum likelihood estimators, in Section 4.2. Section 4.3, three applications in real data sets to illustrate the proposed methodology are presented, in which one of them is a newly added data set to the literature, related to a study about breast cancer in the A C.Camargo Cancer Center, São Paulo, Brazil. Finally, in Section 4.4 the conclusions of this chapter are presented.

4.1 Defective models induced by frailty term

In this section the defective models using a frailty term are defined. A problem in frailty models is the choice of the distribution of the random effect. The frailty distribution most often applied is the gamma distribution Clayton (1978), Vaupel, Manton and Stallard (1979), Missov (2010) and Missov (2013). However, other choices can be considered, such as the positive stable distribution (HOUGAARD, 1986b), a three-parameter distribution (PVF) (HOUGAARD, 1986a), the compound Poisson distribution (AALEN, 1992), the log-normal and inverse Gaussian distributions (TOMAZELLA, 2003), among others.

If the distribution of the random effect must have expectation 1, then the model to be identifiable (ELBERS; RIDDER, 1982). In this paper, assume that the random variable V follows a gamma distribution with shape parameter k and inverse scale parameter, λ ($V \sim \Gamma(k,\lambda)$), with $\mathbb{E}(V) = k/\lambda$ and $\mathrm{Var}(V) = k/\lambda^2$ (WIENKE, 2003). Here, $\mathbb{E}(V) = k/\lambda = 1$ and $\mathrm{Var}(V) = k/\lambda^2 = \theta$ are considered, being $k = \lambda = \theta^{-1}$.

Suppose that $F_0(t)$ is the cumulative function distribution, where $S_0(t)$ is a proper or a not proper survival function, $h_0(t)$ is the respective hazard function and consider a nonnegative unobservable random variable V that denote the frailty term. This way, the hazard model with a frailty term is given by

$$h(t|V) = V\,h_0(t).$$

The conditional survival function is obtained as follows

$$S(t) = \mathbb{E}[S(t|V)] = \mathbb{E}\left[e^{-H(t|V)}\right] = \mathbb{E}\left[e^{-V H_0(t)}\right] = \mathscr{L}_g\left[H_0(t)\right],$$

where $\mathscr{L}_g[\cdot]$ denotes the Laplace transform of frailty distribution.

The Laplace transform of the gamma frailty distribution is expressed by

$$\mathscr{L}_g(s) = (1+\theta s)^{-1/\theta}.$$

The unconditional survival, density and hazard functions in the gamma frailty model are expressed, respectively

$$S(t) = \left[1-\theta \log S_0(t)\right]^{-1/\theta},$$

$$f(t) = h_0(t)\left[1-\theta \log S_0(t)\right]^{-1-1/\theta},$$

$$h(t) = h_0(t)\left\{1-\theta \log S_0(t)\right\}^{-1},$$

where $S_0(t)$ is a proper or not proper survival function.

It is worth to mentioning, that in Equation (4.1) if $S_0(t)$ is not proper survival function then $S(t)$ is also not proper. Notice that, when $S_0(t)$ is a not proper, the survival function in (4.1) is positive if, and only if, $\theta \log[S_0(t)] < 1$. The main result of this chapter is that if $S_0(t)$ is defective then $S(t)$ is also defective, as stated in below.

Theorem 2. *If $S_0(t)$ is a survival function of a defective distribution and $\theta \log[S_0(t)] < 1$, then $S(t) = \left[1-\theta \log S_0(t)\right]^{-1/\theta}$ is also a survival function from a defective distribution.*

Proof. Suppose the limit of $S_0(t)$ is equal to $p_0 \in (0,1)$. Then

$$\lim_{t\to\infty} S(t) = \lim_{t\to\infty}\left[1-\theta \log S_0(t)\right]^{-1/\theta} = \left[1-\theta \log p_0\right]^{-1/\theta}.$$

If $p_0 \in (0,1)$ then $\log p_0 < 0$. Suppose that $\theta > 0$ then

$$\theta \log p_0 < 0 \Leftrightarrow -\theta \log p_0 > 0 \Leftrightarrow 1 - \theta \log p_0 > 1 \Leftrightarrow (1 - \theta \log p_0)^{-1/\theta} < 1.$$

Therefore, if $\theta \log[S_0(t)] < 1$, we have

$$\lim_{t\to\infty} S(t) = [1 - \theta \log p_0]^{-1/\theta} \in (0,1).$$

The proof is complete. □ □

The proof is similar when $\theta < 0$.

When $S(t)$ is a defective distribution, the proportion of immunity in the population is calculated as the limit of the survival function

$$p = \lim_{t\to\infty} S(t).$$

When the data has a cure fraction, the defective models with frailty terms can result in negatively valued estimates of the parameter θ. Figure 12 presents the cure fraction of the Gompertz distribution (left graphics) and inverse Gaussian distribution (right graphics) for different values of θ. Note that depending on the cure fraction values, the parameter θ can assume negative values, violating the main assumption of the frailty models. So the parameter θ can not be interpret as a frailty term, when it has a defective parameter. This is a question which is not commented in Kettunen *et al.* (1991), which considers an estimate of the frailty term even estimating the defective parameter negative.

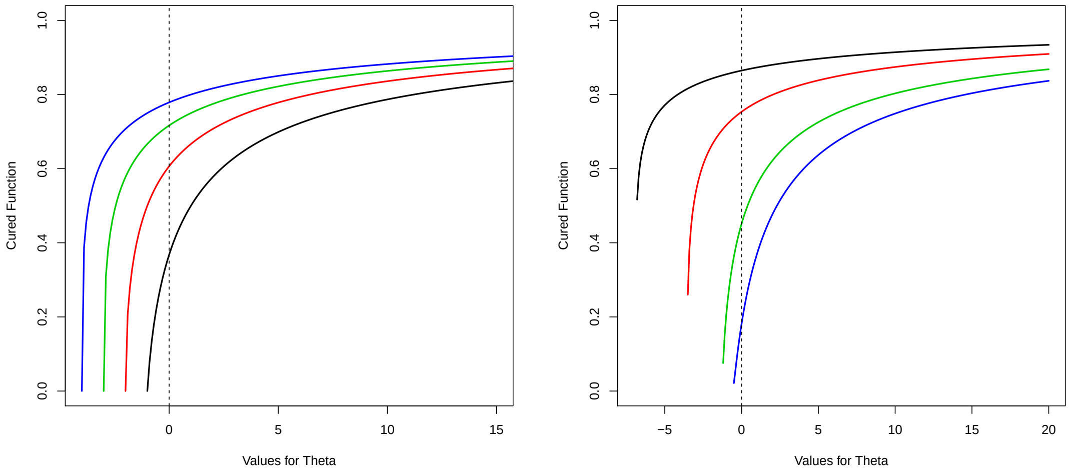

Figure 12 – Graphics of the cure fraction of the Gompertz and inverse Gaussian varying the frailty parameter θ.

In the next section, the defective Gompertz and the defective Inverse Gaussian models with a gamma frailty term will be defined.

4.1.1 Defective gamma-Gompertz model

The defective Gompertz model induced by a frailty term is defined considering that the lifetime of the individuals have a Gompertz distribution with parameter $a < 0$, $b > 0$.

The survival function of the Gompertz distribution with gamma frailty term are defined as

$$S(t) \;=\; [1-\theta \log S_0(t)]^{-1/\theta} = \left\{1+\frac{\theta b\,(e^{at}-1)}{a}\right\}^{-1/\theta}, \tag{4.1}$$

where $a > 0$, $b > 0$ and $\theta > 0$. When $a < 0$, it has the defective gamma-Gompertz distribution and $\theta \in \mathbb{R}^*$.

Its density function is given by

$$f(t) \;=\; h_0(t)\,[1-\theta \log S_0(t)]^{-1-1/\theta} = be^{at}\left\{1+\frac{\theta b\,(e^{at}-1)}{a}\right\}^{-1-1/\theta}.$$

The hazard function is

$$h(t) \;=\; h_0(t)\,[1-\theta \log S_0(t)]^{-1} = be^{at}\left\{1+\frac{\theta b\,(e^{at}-1)}{a}\right\}^{-1}.$$

Theorem 3. *The gamma-Gompertz distribution is defective when $a < 0$.*

Proof. If

$$\lim_{t\to\infty} S(t) = \left\{1-\theta \lim_{t\to\infty}[\log\ S_0(t)]\right\}^{-1/\theta} = \left\{1+\theta \lim_{t\to\infty}\left[\frac{b\,(e^{at}-1)}{a}\right]\right\}^{-1/\theta} = \left\{1-\frac{b\theta}{a}\right\}^{-1/\theta}.$$

Suppose that $a < 0$ and $\theta > 0$ then $-\frac{b}{a} > 0$. So,

$$-\frac{\theta b}{a} > 0 \Leftrightarrow 1-\frac{\theta b}{a} > 1 \Leftrightarrow \left\{1-\frac{\theta b}{a}\right\}^{\frac{1}{\theta}} > 1^{\frac{1}{\theta}} \Leftrightarrow \left\{1-\frac{\theta b}{a}\right\}^{-\frac{1}{\theta}} < 1.$$

Now suppose that $a < 0$ and $\theta < 0$, then $-\frac{b}{a} > 0$. So,

$$-\frac{\theta b}{a} < 0 \Leftrightarrow 1-\frac{\theta b}{a} < 1 \Leftrightarrow \left\{1-\frac{\theta b}{a}\right\}^{\frac{1}{\theta}} > 1^{\frac{1}{\theta}} \Leftrightarrow \left\{1-\frac{\theta b}{a}\right\}^{-\frac{1}{\theta}} < 1.$$

If $\theta \log S_0(t) < 1$, we have

$$\lim_{t\to\infty} S(t) = \left\{1-\frac{\theta b}{a}\right\}^{-\frac{1}{\theta}} \in (0,1).$$

The proof is similar to the proof of Theorem 2. □ □

Note that to $a > 0$, $\lim\limits_{t\to\infty} S(t) = 0$, then the Equation (4.1) is proper and it has the gamma-Gompertz model. If $a < 0$, $\lim\limits_{t\to\infty} S(t) \in (0,1)$ then the Equation (4.1) is not proper and we have the defective gamma-Gompertz model. Then, when $a < 0$ we have a cure rate model, with cure fraction p estimated by

$$p = \left(1 - \frac{\theta b}{a}\right)^{-\frac{1}{\theta}}.$$

4.1.2 Defective gamma-inverse Gaussian model

The defective inverse Gaussian model induced by the gamma frailty term is defined considering that the lifetime of the individuals have inverse Gaussian distribution with parameters $a < 0$, $b > 0$. This model is referred as defective gamma-inverse Gaussian model. The survival function of the gamma-inverse Gaussian model is given by

$$S(t) = [1-\theta\log S_0(t)]^{-1/\theta} = \left\{1-\theta\log\left\{1-\left[\Phi\left(\tfrac{-1+at}{\sqrt{bt}}\right)+e^{\frac{2a}{b}}\Phi\left(\tfrac{-1-at}{\sqrt{bt}}\right)\right]\right\}\right\}^{-1/\theta},$$

where $a > 0$, $b > 0$ and $\theta > 0$. When $a < 0$, we have the defective gamma-inverse Gaussian model and $\theta \in \mathbb{R}^*$.

The density function is given by

$$f(t) = \frac{\frac{1}{\sqrt{2b\pi t^3}}\exp\left\{-\frac{1}{2bt}(1-at)^2\right\}}{1-\left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right]}\left\{1-\theta\log\left[1-\left[\Phi\left(\tfrac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\tfrac{-1-at}{\sqrt{bt}}\right)\right]\right]\right\}^{-1-1/\theta}.$$

The hazard function is defined as

$$h(t) = \frac{\frac{1}{\sqrt{2b\pi t^3}}\exp\left\{-\frac{1}{2bt}(1-at)^2\right\}}{1-\left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right]}\left\{1-\theta\log\left[1-\left[\Phi\left(\tfrac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\tfrac{-1-at}{\sqrt{bt}}\right)\right]\right]\right\}^{-1}.$$

Theorem 4. *The gamma-inverse Gaussian distribution is defective when $a < 0$.*

Proof.

$$\begin{aligned}\lim_{t\to\infty} S(t) &= \left\{1-\theta\lim_{t\to\infty}[\log S_0(t)]\right\}^{-\frac{1}{\theta}}\\ &= \left\{1+\theta\lim_{t\to\infty}\left[\log\left(1-\left(\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{\frac{2a}{b}}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right)\right)\right]\right\}^{-\frac{1}{\theta}}\\ &= \left\{1-\left[\theta\log\left\{1-e^{\frac{2a}{b}}\right\}\right]\right\}^{-\frac{1}{\theta}}.\end{aligned}$$

Suppose $a < 0$ and $\theta > 0$, then $\left(1-e^{2a/b}\right) \in (0,1)$

$$\begin{aligned}&\log\left(1-e^{2a/b}\right) < 0 \;\Leftrightarrow\; -\log\left(1-e^{2a/b}\right) > 0 \;\Leftrightarrow\\ &-\theta\log\left(1-e^{2a/b}\right) > 0 \Leftrightarrow\\ &\left\{1-\theta\log\left(1-e^{\frac{2a}{b}}\right)\right\} > 1 \;\Leftrightarrow\; \left\{1-\theta\log\left(1-e^{\frac{2a}{b}}\right)\right\}^{-1/\theta} < 1.\end{aligned}$$

Now, suppose $a < 0$ and $\theta < 0$, then

$$\begin{aligned} \log\left(1-e^{2a/b}\right) < 0 \;\; &\Leftrightarrow \;\; -\log\left(1-e^{2a/b}\right) > 0 \Leftrightarrow \\ \left\{-\theta\log\left(1-e^{\frac{2a}{b}}\right)\right\} < 0 \;\; &\Leftrightarrow \;\; \left\{1-\theta\log\left(1-e^{\frac{2a}{b}}\right)\right\} < 1 \Leftrightarrow \\ \left\{1-\theta\log\left(1-e^{\frac{2a}{b}}\right)\right\}^{1/\theta} > 1 \;\; &\Leftrightarrow \;\; \left\{1-\theta\log\left(1-e^{\frac{2a}{b}}\right)\right\}^{-1/\theta} < 1. \end{aligned}$$

Therefore, if $\theta \log S_0(t) < 1$, we have

$$\lim_{t\to\infty} S(t) = \left\{1-\theta\log\left(1-e^{\frac{2a}{b}}\right)\right\}^{-\frac{1}{\theta}} \in (0,1).$$

The proof is complete. □ □

Note that if $a > 0$, $\lim\limits_{t\to\infty} S(t) = 0$, then the Equation (4.2) is proper and we have the gamma-inverse Gaussian frailty model. When $a < 0$ the Equation (4.2) is improper and we have the defective gamma-inverse Gaussian model. We calculate the cure fraction p for the defective gamma-inverse Gaussian model as

$$p = \left\{1-\theta\log\left(1-e^{\frac{2a}{b}}\right)\right\}^{-\frac{1}{\theta}}.$$

4.1.3 Defective regression models induced by a frailty term

In this section, a defective model induced by a frailty term with covariate information is proposed. Consider $\boldsymbol{x}^\top = (1, x_1, \ldots, x_p)$ a vector of covariates and $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \ldots, \beta_p)$ a vector of regression coefficients. Given a cumulative function distribution $F_0(t)$, with survival function $S_0(t)$ and $h_0(t)$ its hazard function, the corresponding conditional hazard function of the regression model with a frailty term is

$$h(t|V,\boldsymbol{x}) = V\, h_0(t) e^{\boldsymbol{x}^\top\boldsymbol{\beta}},$$

where $V \sim \Gamma(\theta^{-1}, \theta^{-1})$.

Then, the survival function of the regression model with a gamma frailty term is given by

$$S(t|\boldsymbol{x}) \;\; = \;\; \left[1-\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}} \log S_0(t)\right]^{-1/\theta}.$$

When $S_0(t)$ is a defective distribution, we have a defective regression models induced by a frailty term. Notice that if $S_0(t)$ is not proper, the expression above is positive if, and only if, $\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}} \log S_0(t) < 1$.

The density and hazard functions are, respectively

$$f(t|\boldsymbol{x}) \;\; = \;\; h_0(t) e^{\boldsymbol{x}^\top\boldsymbol{\beta}} \left[1-\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}} \log S_0(t)\right]^{-1-1/\theta}$$

and
$$h(t|\boldsymbol{x}) \;=\; h_0(t)e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\left[1-\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\log S_0(t)\right]^{-1}.$$

In the next section, the defective regression model to the gamma-Gompertz and gamma-inverse Gaussian distributions will be introduced.

4.1.3.1 Defective gamma-Gompertz regression model

Here the defective Gompertz model with frailty term including covariate information is defined. The survival function of the gamma-Gompertz regression model is defined as

$$S(t|\boldsymbol{x}) = \left\{1-\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\log S_0(t)\right\}^{-1/\theta} = \left\{1+\frac{\theta b e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\left(e^{at}-1\right)}{a}\right\}^{-1/\theta},$$

where $a>0$, $b>0$ and $\theta>0$. However, the gamma-Gompertz has an identifiability problem between the parameters b and e^{β_0}. To overcome this problem, $e^{\beta_0^*} = be^{\beta_0} = e^{\log b}e^{\beta_0} = e^{\log b+\beta_0}$ is used, which is the same as setting $b=1$. Not, that use this transformation are not losing any information about the parameter b, as the same information is being considered in $e^{\beta_0^*}$. Then, the survival function of the gamma-Gompertz model is

$$S(t|\boldsymbol{x}) = \left\{1+\frac{\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\left(e^{at}-1\right)}{a}\right\}^{-1/\theta}.$$

The density function of the gamma-Gompertz model induced by a frailty term is given by

$$f(t|\boldsymbol{x}) = h_0(t)e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\left\{1-\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\log\left[S_0(t)\right]\right\}^{-1-1/\theta} = e^{at+\boldsymbol{x}^\top\boldsymbol{\beta}}\left\{1+\frac{\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\left(e^{at}-1\right)}{a}\right\}^{-1-1/\theta}.$$

The hazard function of the gamma-Gompertz is given by

$$h(t|\boldsymbol{x}) = h_0(t)e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\left[1-\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\log S_0(t)\right]^{-1} = e^{at+\boldsymbol{x}^\top\boldsymbol{\beta}}\left\{1+\frac{\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\left(e^{at}-1\right)}{a}\right\}^{-1}.$$

If $a<0$, we have the defective gamma-Gompertz regression model and $\theta\in\mathbb{R}^*$. The cure fraction p is

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

4.1.3.2 Defective gamma-inverse Gaussian regression model

Here, the gamma-inverse Gaussian model with covariate information is defined. The survival function of the gamma-inverse Gaussian regression model is given by

$$\begin{aligned} S(t|\boldsymbol{x}) &= \left[1-\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\log S_0(t)\right]^{-1/\theta} \\ &= \left\{1-\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\log\left\{1-\left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{\frac{2a}{b}}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right]\right\}\right\}^{-1/\theta}, \end{aligned}$$

where $a > 0$, $b > 0$ and $\theta > 0$.

The density and hazard functions are, respectively

$$
\begin{aligned}
f(t|\boldsymbol{x}) \;=\; & \frac{\frac{1}{\sqrt{2b\pi t^3}}\exp\left\{-\frac{1}{2bt}\left(1-at\right)^2\right\}}{1-\left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right]}e^{\boldsymbol{x}^\top\boldsymbol{\beta}} \\
& \left\{1-\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\log\left[1-\left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right]\right]\right\}^{-1-1/\theta},
\end{aligned}
$$

and

$$
\begin{aligned}
h(t|\boldsymbol{x}) \;=\; & \frac{\frac{1}{\sqrt{2b\pi t^3}}\exp\left\{-\frac{1}{2bt}\left(1-at\right)^2\right\}}{1-\left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right]}e^{\boldsymbol{x}^\top\boldsymbol{\beta}} \\
& \left\{1-\theta e^{\boldsymbol{x}^\top\boldsymbol{\beta}}\log\left[1-\left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right]\right]\right\}^{-1}.
\end{aligned}
$$

When $a < 0$, we have the defective gamma-inverse Gaussian regression model and $\theta \in \mathbb{R}^*$. In this case the cure fraction p is calculated by

$$p = \left\{1-\theta e^{\boldsymbol{x}\boldsymbol{\beta}}\log\left(1-e^{\frac{2a}{b}}\right)\right\}^{-\frac{1}{\theta}}.$$

In the next section, the inference for the defective gamma-Gompertz and gamma-inverse Gaussian models will be presented.

4.1.4 Inference

Here, the estimation procedures regards the defective gamma-Gompertz model and the defective gamma-inverse Gaussian model are discussed.

Consider a sample of size n and the observed time is $T_i = \min\left(W_i, C_i\right)$ where W_i is the time failure, C_i the censoring time and let δ_i the failure indicator, that is, $\delta_i = 1$ if $T_i = W_i$ and $\delta_i = 0$ otherwise, for $i = 1,\ldots,n$. The observed data are represent by $D = (\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{X})$, $\boldsymbol{t} = (t_1,\ldots,t_n)^\top$, $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_n)^\top$ and $\boldsymbol{X}$ is a $n\times p$ matrix containing the covariates. Suppose that the data is independently and identically distributed and come from a distribution with density and survival functions specified by $f(\cdot,\boldsymbol{\theta})$ e $S(\cdot,\boldsymbol{\theta})$, where $\boldsymbol{\theta} = \left(a,b,\boldsymbol{\beta}^\top\right)^\top$ denoting a vector of parameters.

The likelihood function of $\boldsymbol{\theta}$ can be written as

$$L(\boldsymbol{\theta},D) \propto \prod_{i=1}^{n} f\left(t_i,\boldsymbol{\theta}\right)^{\delta_i} S\left(t_i,\boldsymbol{\theta}\right)^{(1-\delta_i)}.$$

The corresponding log-likelihood function is

$$\log L(\boldsymbol{\theta},D) \;=\; const + \sum_{i=1}^{n}\delta_i \log f\left(t_i,\boldsymbol{\theta}\right) + \sum_{i=1}^{n}\left(1-\delta_i\right)\log S\left(t_i,\boldsymbol{\theta}\right).$$

For the gamma-Gompertz distribution the log-likelihood function for $\boldsymbol{\theta}$ is

$$\begin{aligned}\log L\left(\boldsymbol{\theta},D\right) &= const+\sum_{i=1}^{n}\delta_i\log f(t_i,\boldsymbol{\theta})+\sum_{i=1}^{n}(1-\delta_i)\log S(t_i,\boldsymbol{\theta}) \qquad (4.2)\\ &= const+\sum_{i=1}^{n}\delta_i\boldsymbol{x}_i^{\top}\boldsymbol{\beta}+a\sum_{i=1}^{n}\delta_i t_i-\left(\frac{1}{\theta}+1\right)\sum_{i=1}^{n}\delta_i\log\left\{1+\frac{\theta e^{\boldsymbol{x}_i^{\top}\boldsymbol{\beta}}(e^{at_i}-1)}{a}\right\}\\ &\quad -\frac{1}{\theta}\sum_{i=1}^{n}(1-\delta_i)\log\left\{1+\frac{\theta e^{\boldsymbol{x}_i^{\top}\boldsymbol{\beta}}(e^{at_i}-1)}{a}\right\}.\end{aligned}$$

For the gamma-inverse Gaussian distribution the log-likelihood function for $\boldsymbol{\theta}$ is

$$\begin{aligned}\log L\left(\boldsymbol{\theta},D\right) &= const+\sum_{i=1}^{n}\delta_i\log f(t_i,\boldsymbol{\theta})+\sum_{i=1}^{n}(1-\delta_i)\log S(t_i,\boldsymbol{\theta}) \qquad (4.3)\\ &= const+\sum_{i=1}^{n}\delta_i\log\left(\frac{\frac{1}{\sqrt{2b\pi t_i^3}}\exp\left\{-\frac{1}{2bt}\left(1-at_i\right)^2\right\}}{1-\left[\Phi\left(\frac{-1+at_i}{\sqrt{bt_i}}\right)+e^{2a/b}\Phi\left(\frac{-1-at_i}{\sqrt{bt_i}}\right)\right]}\right)+\sum_{i=1}^{n}\boldsymbol{\delta_i x_i^{\top}\beta}\\ &\quad -\left(1+\frac{1}{\theta}\right)\sum_{i=1}^{n}\delta_i\left\{1-\theta e^{\boldsymbol{x}_i^{\top}\boldsymbol{\beta}}\log\left[1-\left(\Phi\left(\frac{-1+at_i}{\sqrt{bt_i}}\right)+e^{\frac{2a}{b}}\Phi\left(\frac{-1-at_i}{\sqrt{bt_i}}\,)\right)\right]\right\}\\ &\quad -\frac{1}{\theta}\sum_{i=1}^{n}(1-\delta_i)\log\left\{1-\theta e^{\boldsymbol{x}_i^{\top}\boldsymbol{\beta}}\log\left[1-\left(\Phi\left(\frac{-1+at_i}{\sqrt{bt_i}}\right)+e^{\frac{2a}{b}}\Phi\left(\frac{-1-at_i}{\sqrt{bt_i}}\right)\right)\right]\right\}.\end{aligned}$$

The log-likelihood functions (4.2) and (4.3), can be maximized numerically to obtain the maximum likelihood estimates. There are various routines available for numerical maximization. Here, the package optim of the R software for the numerical maximization are used (R Core Team, 2013). And the "BFGS" method is used for maximization, see the manual of the optim package for more details. Confidence intervals for the parameters were based on asymptotic normality. If $\widehat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimator of $\boldsymbol{\theta}$ then it is well known that the distribution of $\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}$ can be approximated by a q-variate normal distribution with zero means and covariance matrix $I^{-1}\left(\widehat{\boldsymbol{\theta}}\right)$, where $I\left(\boldsymbol{\theta}\right)$ denotes the observed information matrix. So, an approximate $100(1-\alpha)$ percent confidence interval for θ_i is $\left(\widehat{\theta}_i-z_{\alpha/2}\sqrt{I^{ii}},\widehat{\theta}_i+z_{\alpha/2}\sqrt{I^{ii}}\right)$, where I^{ii} denotes the ith diagonal element of the inverse of I evaluated at $\widehat{\boldsymbol{\theta}}$ and z_a denotes the $100(1-a)$ percentile of the standard normal random variable.

In the next section, simulation studies were performed to check the asymptotes of the maximum likelihood estimates.

4.2 Simulation studies

In this section, a study with simulated data is proposed to check the properties of the maximum likelihood estimator for the proposed models.

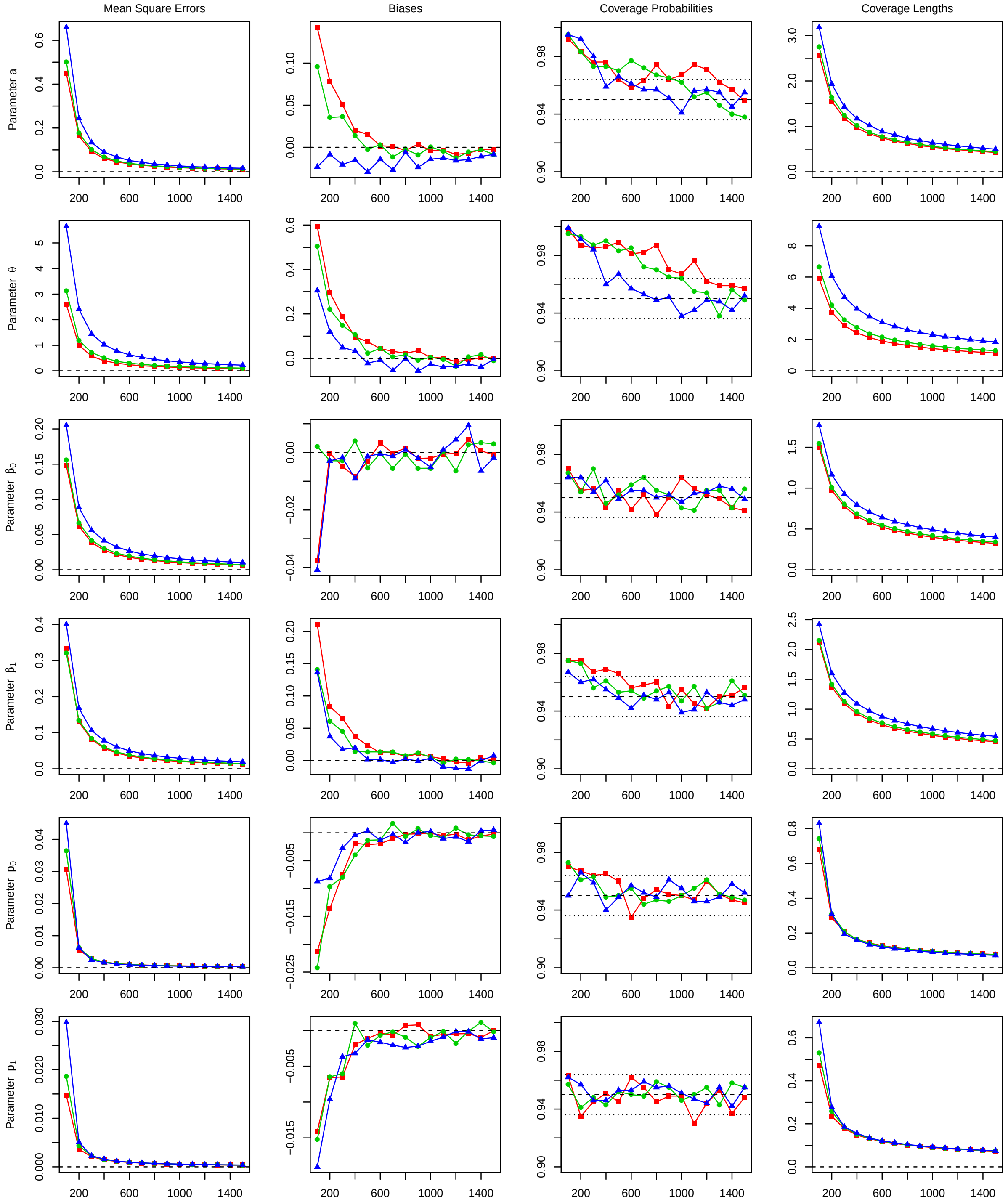

Figure 13 – Mean squared errors, bias, coverage probabilities and covarage length of $(\hat{a},\hat{\theta},\hat{\beta}_0,\hat{\beta}_1,\hat{p_0},\hat{p_1})$ versus n for the defective gamma-Gompertz model. The red, green and blue lines (squares, circles and triangles) represents the scenario where $\theta = 0.75, 1, 2$, respectively.

Using the algorithm describe in Section 2.8, six simulations to check the maximum likelihood estimates are proposed, with three scenarios for the defective gamma-Gompertz model and three for the defective gamma-inverse Gaussian model.

The parameters used in the defective gamma-Gompertz distribution were $(a,\theta,\beta_0,\beta_1) =$

$(-1, \theta, -0.5, 1)$, with $\theta = 0.75, 1, 2$. Thus, the cure fractions p_0 and p_1 are 0.6722 and 0.4824 for $\theta = 0.75$, 0.6225 and 0.3775 for $\theta = 1$ and 0.6722 and 0.4824 for $\theta = 2$. The parameters used in the defective gamma-inverse Gaussian distribution were $(a, b, \theta, \beta_0, \beta_1) = (-1, 2, \theta, 1, 1)$, also with $\theta = 0.75, 1, 2$. Thus, the cure fractions p_0 and p_1 are 0.4147 and 0.1852 for $\theta = 0.75$, 0.4451 and 0.2278 for $\theta = 1$ and 0.5350 and 0.3586 for $\theta = 2$.

In each scenario, for each parameter the mean square error, the bias, the coverage probability and range of the confidence interval for different sample sizes were calculated. The sample sizes were considered $n = 100, 200, ..., 1.500$. The delta method was used to estimate the variance of cure fractions. All calculations were developed in R software. Figures 13 and 14 illustrate the results obtained by the defective gamma-Gompertz and defective gamma-inverse Gaussian models, respectively. The red curves are for the choice of $\theta = 0.75$, the green curves for $\theta = 1$ and blue curves for $\theta = 2$.

From these simulations we can conclude the following: i) the mean square error decreases and approaches zero as the sample size increases; ii) the mean square error is smaller for the lowest values of θ; iii) the mean square error of the cure fractions is generally lower than for the other parameters; iv) all parameters in all cases converge to their actual values that is, the maximum likelihood estimator of these models is asymptotically not biased; v) the bias of cure fractions are very small, even in the smallest sample values;vi) the coverage probabilities are around 0.95 for the parameters a, b, θ, β_0 , β_1, p_0 and p_1; vii) the ranges of confidence intervals decrease as the sample size increases; viii) in general, when the amplitude of the interval is smaller than the other, the smaller the value of θ will be.

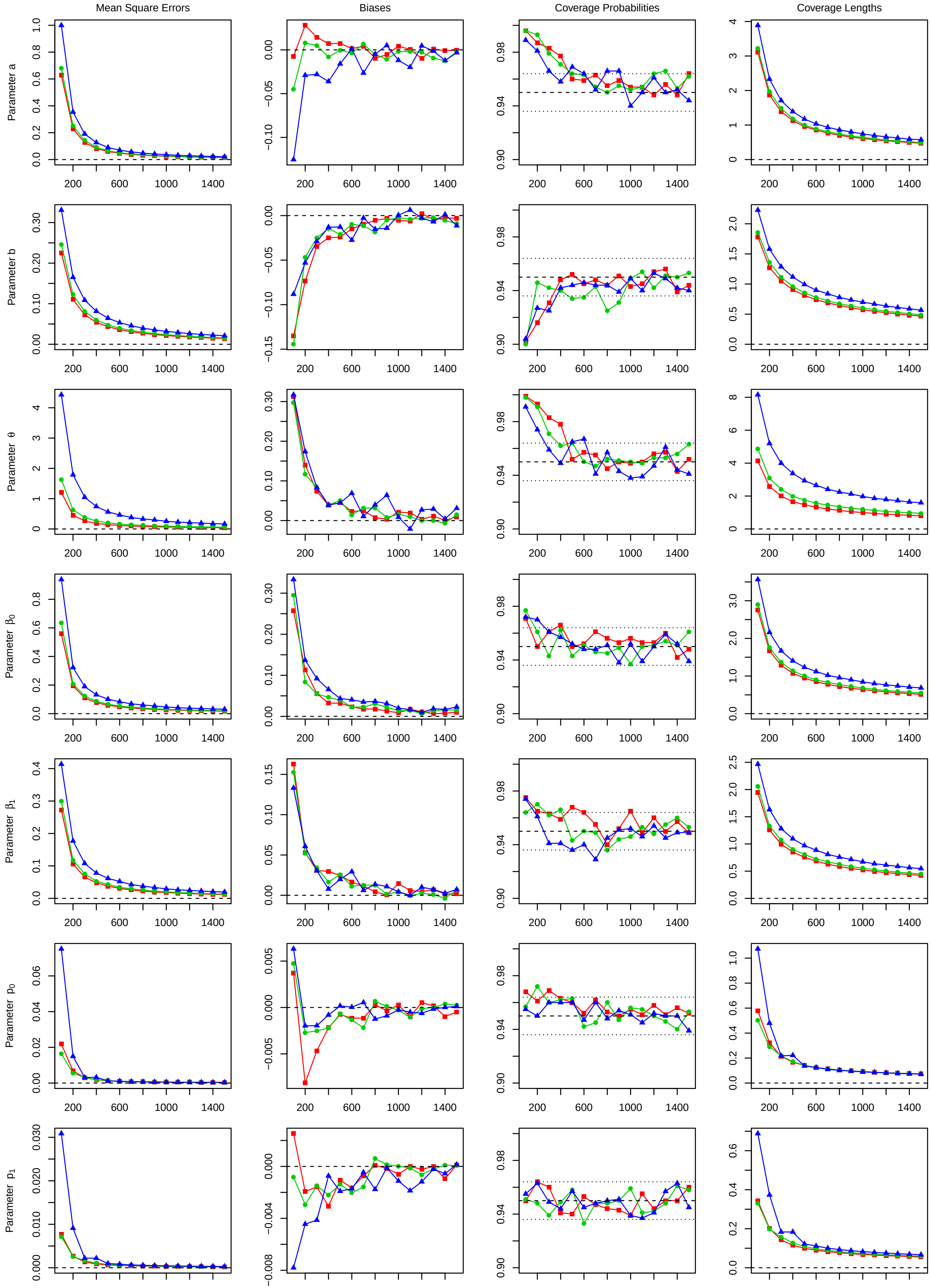

Figure 14 – Mean squared errors, bias, coverage probabilities and covarage length of $(\hat{a}, \hat{b}, \hat{\theta}, \hat{\beta}_0, \hat{\beta}_1, \hat{p_0}, \hat{p_1})$ versus n for the the defective gamma-inverse Gaussian model. The red, green and blue lines (squares, circles and triangles) represents the scenario where $\theta = 0.75, 1, 2$, respectively.

The next section illustrates the proposed models in four real data sets.

4.3 Application

In this section some applications of the proposed models were discussed. Four real data sets were used, the same data set presented in Section 3.5. The fit of the defective models induced by frailty term were compared with the frailty mixture models (PRICE; MANATUNGA, 2001). We used of the AIC, BIC as comparative measures.

The gamma frailty mixture model proposed by Price and Manatunga (2001) is defined as

$$S(t|\boldsymbol{x}) = p(\boldsymbol{x}) + (1 - p(\boldsymbol{x}))\,(1 - \theta log(S_0(t)))^{1/\theta}$$

where $\theta > 0$, $S_0(t)$ is the Gompertz distribution or inverse-Gaussian and $p(\boldsymbol{x}) = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}$.

The plot of the Kaplan-Meier non-parametric estimator curve with the survival curve of the models proposed were presented. As close to the parametric models gets to the Kaplan-Meier curve, better the fit is. In all applications the R software was used, with the optimization procedure given by the function *optim*. The delta method was used to calculate the standard error of the cure rates, both in the frailty mixture model and the defective model induced by frailty term.

4.3.1 Colon cancer

Here colon data set is considered, this data set was presented in Chapter 3. Details of this data set can be found in Laurie *et al.* (1989).

Table 6 – Maximum likelihood estimates, standard error (SE), 95% confidence intervals (CIs), AIC and BIC criteria according to gamma-Gompertz and frailty mixture gamma-Gompertz for the colon cancer data.

	Defective gamma-Gompertz				Mixture gamma-Gompertz			
Parameter	Estimate	Std. Error	Lower 95% CI	Upper 95% CI	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
a	-4.343	0.667	-5.651	-3.035	90.322	24.848	41.621	139.024
b	-	-	-	-	0.544	0.221	0.110	0.978
θ	-1.775	0.566	-2.885	-0.666	22.178	6.571	9.300	35.057
β_0	0.574	0.051	0.473	0.674	-0.063	0.058	-0.176	0.050
β_1	0.147	0.060	0.029	0.265	-0.523	0.150	-0.816	-0.229
p_0	0.483	0.016	0.452	0.514	0.484	0.014	0.456	0.512
p_1	0.356	0.037	0.285	0.428	0.358	0.033	0.294	0.422
AIC	1498.70				1443.69			
BIC	1520.81				1471.33			

Table 6 summarizes the results from the defective gamma-Gompertz model and the frailty mixture gamma-Gompertz model in the colon cancer data. Notice that $\hat{a} < 0$ in the defective gamma-Gompertz, which indicates cure rate, although the parameter θ has not interpretation as a frailty term. The both models a reasonable fit and presents a quite close estimate for the proportion of cured. Both models provide cured fraction estimates of $p_0 = 0.48$ and $p_1 = 0.36$.

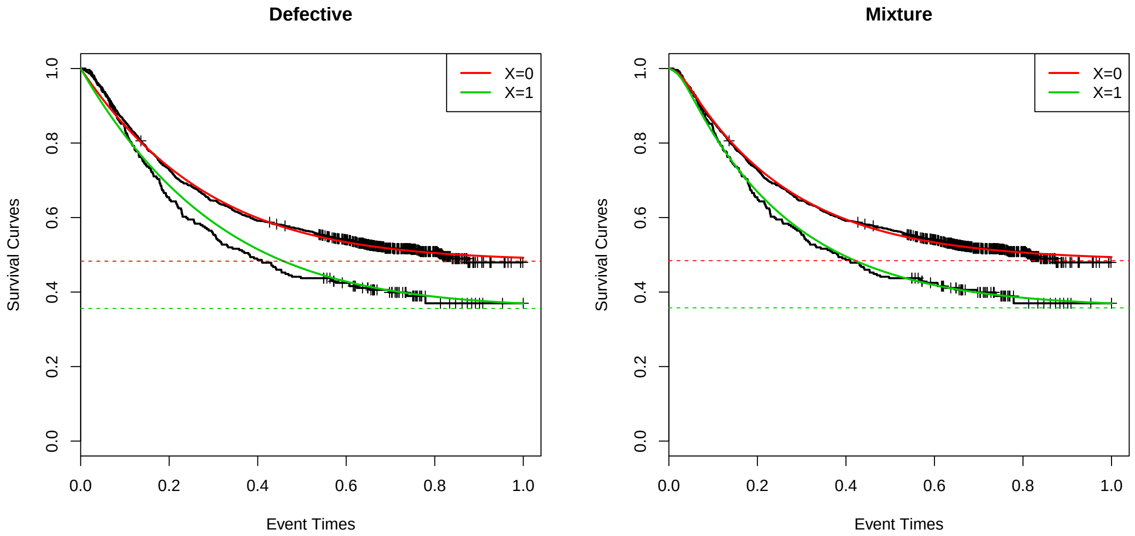

Figure 15 – Estimated survival curves of the gamma-Gompertz model and frailty mixture gamma-Gompertz for the colon data set.

This means that patients that present adherence in the nearby organs have a smaller estimated cure fraction. The difference in the estimated cure fraction between the two groups is around 0.12. The AIC and BIC measures to the frailty mixture model is a bit smaller than the defective model.

4.3.2 Divorce data

The second data set considered is the divorce data set. The Figure 16 illustrates the estimated survival curves for the gamma-Gompertz model and frailty mixture gamma-Gompertz model. We can see the both models have a reasonable fit, but the frailty mixture model shows a better fit with Kaplan-Meier curve. The AIC and BIC measures to the frailty mixture models is a bit smaller than the defective model.

The Table 7 summarizes the results from the defective gamma-Gompertz and the frailty mixture gamma-Gompertz models. Note that $\hat{a} < 0$ in the defective gamma-Gompertz, which indicates cure rate. Then the parameter θ has not interpretation as a frailty term.

In both models the cure rate is smaller in couple with different ethnicity, i.e., couples with different ethnicity have more odds divorce than couple with same ethnicity. The difference in estimated cure fraction between the two groups is around 0.09.

4.3.3 Breast cancer

This data set was collected at the A.C.Camargo Cancer Center, São Paulo, Brazil. It is a study of patients with breast cancer.

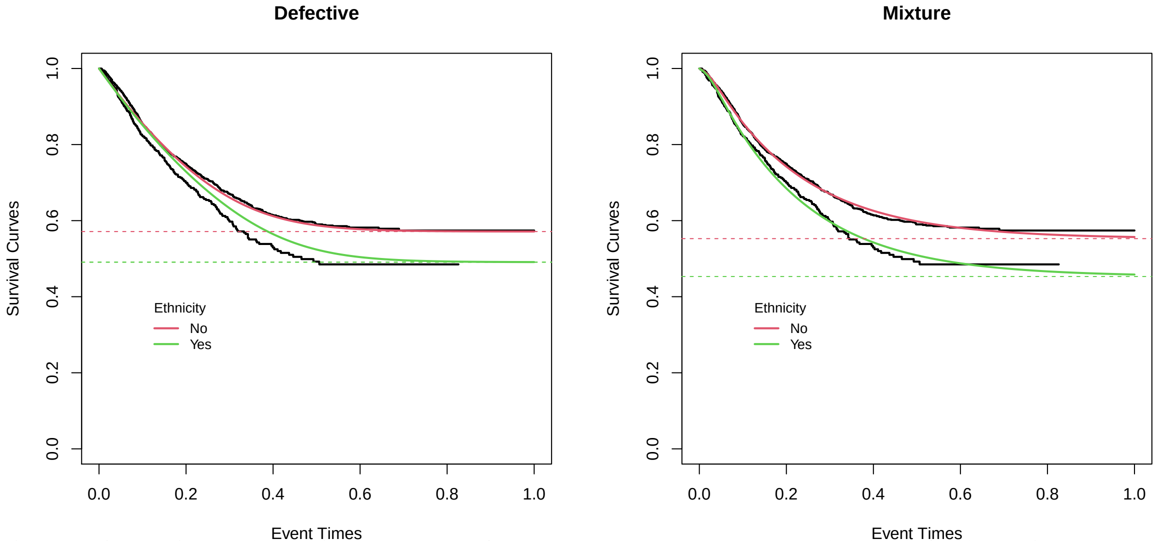

Figure 16 – Estimated survival curves of the gamma-Gompertz model and frailty mixture gamma-Gompertz model with the covariate ethnicity in divorce data.

Table 7 – Maximum likelihood estimates, standard error (SE), 95% confidence intervals (CIs), AIC and BIC criteria according to gamma-Gompertz and frailty mixture gamma-Gompertz for the divorce data.

	Defective gamma-Gompertzl				Mixture gamma-Gompertz			
Parameter	Estimate	SE	Lower 95% CI	Upper 95% CI	Estimate	SE	Lower 95% CI	Upper 95% CI
a	-10.1665	1.7252	-13.5478	-6.7852	77.0297	29.3728	19.4601	134.5993
θ	-6.1765	1.4057	-8.9317	-3.4214	16.2195	6.9238	2.6491	29.7900
β_0	0.4663	0.0487	0.371	0.5617	0.2115	0.0642	0.0858	0.3373
β_1	0.0196	0.016	-0.0117	0.0509	-0.3996	0.1201	-0.6351	-0.1642
b	-	-	-	-	1.0928	0.3584	0.3903	1.7953
p_0	0.5713	0.0132	0.5454	0.5972	0.5527	0.0159	0.5216	0.5838
p_1	0.4908	0.0323	0.4275	0.5541	0.4531	0.0294	0.3955	0.5107
AIC	1467.394				1433.061			
BIC	1491.886				1463.676			

In this application, two models are considered: one model only with the covariate N, because is the only one without missing values and the model with all three covariates, but discarding the observations with missing data, totaling 58 observations. The frailty mixture model is used to compare the defective model just the model only with the covariate N.

Table 8 – Maximum likelihood estimates, standard error (SE), 95% confidence intervals (CIs), AIC and BIC criteria according to gamma-inverse Gaussian and frailty mixture gamma-inverse Gaussian for the breast cancer data. with one covariate (N).

	Defective gamma-inverse Gaussian				Mixture gamma-inverse Gaussian			
Parameter	Estimate	Std. Error	Lower 95% CI	Upper 95% CI	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
a	-5.189	2.588	-10.262	-0.117	5.301	0.955	3.429	7.173
b	1.929	0.659	0.637	3.221	4e-04	0.022	-0.044	0.044
θ	-0.801	4.180	-8.995	7.393	1.551	0.651	0.276	2.827
β_0	3.657	2.596	-1.432	8.746	-1.396	0.715	-2.796	0.005
β_1	1.004	0.930	-0.818	2.826	1.940	0.601	0.763	3.118
p_0	0.825	0.095	0.638	1.000	0.825	0.094	0.641	1.000
p_1	0.538	0.084	0.374	0.703	0.539	0.083	0.377	0.701
AIC	33.908				33.912			
BIC	45.692				45.696			

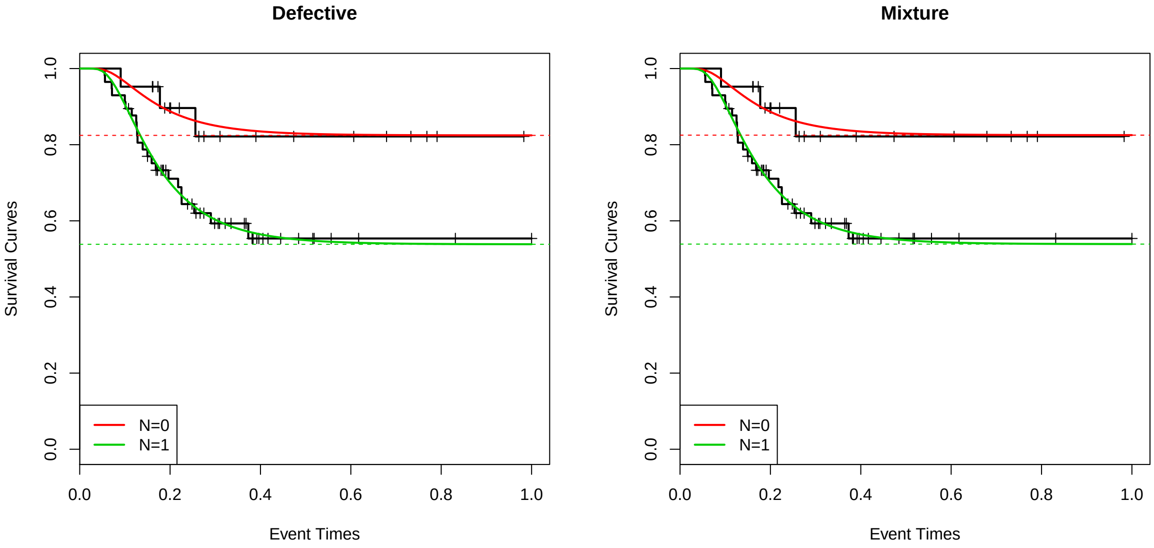

Figure 17 – Estimated survival curves of the gamma-inverse Gaussian model and frailty mixture gamma-inverse Gaussian model with the covariate N.

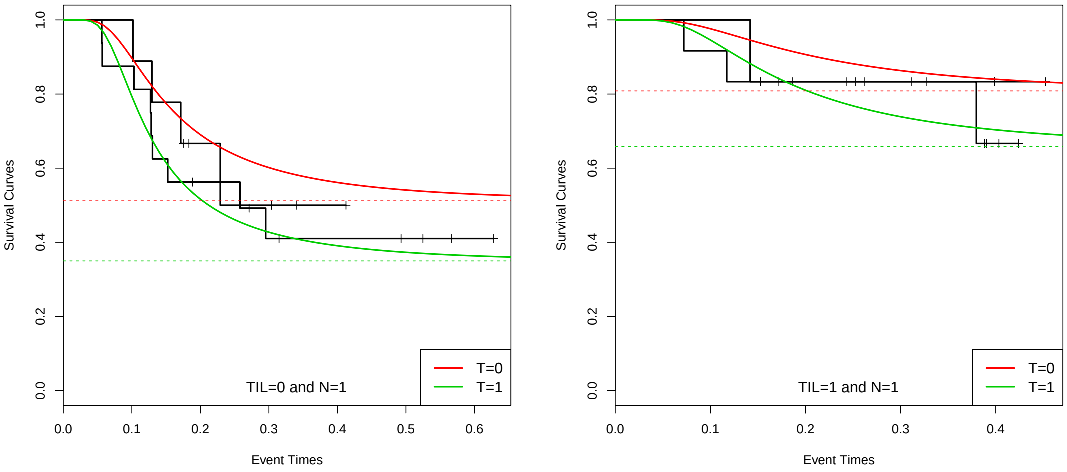

Figure 18 – Estimated survival curves of the gamma-inverse Gaussian model with three covariates. In the left, the survival curves when TIL=0 and N=1. In the right, the survival curves when TIL=1 and N=1.

Table 8 summarizes the results for the fit of gamma-inverse Gaussian model and the frailty mixture gamma-inverse Gaussian in the breast cancer data set with covariate N. Note that $\hat{a} < 0$, so we have a defective gamma-inverse Gaussian model. And the frailty term do not interpretation in this case. The both models have reasonable fit and presents the same estimate for proportion of cured elements estimated around $p_0 = 0.82$ and $p_1 = 0.53$, for the groups $\mathrm{N} = 0$ and $\mathrm{N} = 1$, respectively.

Figure 17 illustrates the estimated survival curves. The survival time of the $\mathrm{N} = 0$ group is higher than the $\mathrm{N} = 1$ group and differs by around 0.29. However, we can not distinguish it with statistical significance. The fit for both models captures the Kaplan-Meier curve very well for both groups. The AIC and BIC measures are very closely in both models. And the both models are appropriate for fit this data with covariate N.

Table 9 shows the maximum likelihood estimates of the gamma-inverse Gaussian model with all three covariates. Notice that $\hat{a} < 0$, so we have a defective model. Table 10 shows all the estimated cure rates, for each scenario in the covariate set. Notice that if $\text{TIL} = 1$ and $\text{N} = \text{T} = 0$, we have the case when the cure rate is higher, 94.7%. When the opposite happens, that is, $\text{TIL} = 0$ and $\text{N} = \text{T} = 1$, the cure rate is $34,97\%$. It was not possible to distinguish any group with statistical significance, mainly because the data set is quite small. Figure 18 present the Kaplan-Meier curves when the covariates TIL and N are fixed equals to 0 and 1, in the left panel, and equals to 1 and 1, in the right panel. We can see that, despite for the small data set, the fitted curves captures reasonably well the respective Kaplan-Meier curves.

Table 9 – Maximum likelihood estimates of the gamma-inverse Gaussian model with three covariates.

Parameter	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
a	-3.4846	2.941	-9.2489	2.2797
b	1.7830	0.6589	0.4916	3.0744
θ	1.7224	2.8664	-3.8956	7.3404
β_0	2.6155	3.272	-3.7976	9.0285
β_1	-1.5809	0.9741	-3.4902	0.3284
β_2	1.5049	1.1275	-0.7049	3.7148
β_3	0.8653	0.9165	-0.9310	2.6616

Table 10 – Estimated cure rates for each configuration of covariates.

Parameter	TIL	N	T	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
p_{000}	0	0	0	0.7972	0.1666	0.4707	0.9999
p_{001}	0	0	1	0.6439	0.2132	0.2262	0.9999
p_{010}	0	1	0	0.5136	0.1663	0.1877	0.8396
p_{011}	0	1	1	0.3497	0.1261	0.1026	0.5968
p_{100}	1	0	0	0.9470	0.0620	0.8256	0.9999
p_{101}	1	0	1	0.8853	0.1133	0.6632	0.9999
p_{110}	1	1	0	0.8084	0.1347	0.5443	0.9999
p_{111}	1	1	1	0.6589	0.1784	0.3093	0.9999

4.3.4 Melanoma

Table 11 summarizes the results for gamma-inverse Gaussian model and frailty mixture gamma-inverse Gaussian model in melanoma data. In the gamma-inverse Gaussian model $a > 0$, so the gamma-inverse Gaussian model is a frailty one and θ is interpretable. Notice that $\hat{\theta} = 1.232$, which indicates a reasonable degree of unobserved heterogeneity in sample. Figure 19 illustrates the estimated survival curves. It is worth mentioning that the fitted models do not captures the Kaplan-Meier curve in every group. This is an visual evidence that this models are not a proper fit for this data set.

Table 11 – Maximum likelihood estimates, standard error (SE), 95% confidence intervals (CIs), AIC and BIC criteria according to gamma-inverse Gaussian and frailty mixture gamma-inverse Gaussian for the melanoma cancer data.

	gamma-inverse Gaussian				Mixture gamma-inverse Gaussian			
Parameter	Estimate	Std. Error	Lower 95% CI	Upper 95% CI	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
a	4.1518	7.025	-9.6169	17.9205	2.1679	0.6789	0.8373	3.4984
b	3.0189	0.4826	2.073	3.9649	0.0002	0.0071	-0.0138	0.0142
θ	1.2322	0.7733	-0.2834	2.7477	0.9849	0.3273	0.3434	1.6265
β_0	-2.3637	1.9947	-6.2733	1.5458	-0.5637	0.1494	-0.8566	-0.2708
β_1	0.5006	0.1185	0.2685	0.7328	2.9825	0.3552	2.2863	3.6788
p_1	-	-	-	-	0.6038	0.0626	0.4812	0.7264
p_2	-	-	-	-	0.4644	0.0678	0.3316	0.5973
p_3	-	-	-	-	0.3304	0.0782	0.1772	0.4837
p_4	-	-	-	-	0.2193	0.0803	0.0618	0.3767
AIC	319.2274				326.8285			
BIC	339.3928				346.9939			

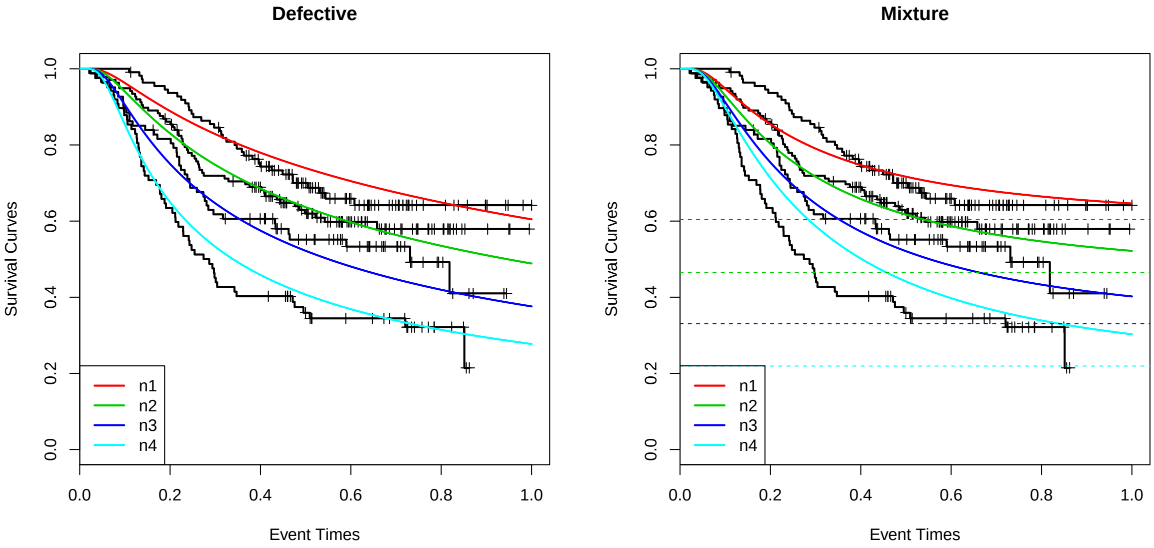

Figure 19 – Estimated survival curves of the gamma-inverse Gaussian model and frailty mixture gamma-inverse Gaussian for the melanoma data set.

According AIC and BIC criteria show that the gamma-inverse Gaussian model is preferable in relation to the frailty mixture gamma-inverse Gaussian model. Notice that the gamma-inverse Gaussian model capture much better the curve for category 4. For category 1 curve, the gamma-inverse Gaussian model capture better the initial times. Then, the gamma-inverse Gaussian model has a better fit than the frailty mixture model and has advantage of it not being necessary to assume the existence of a proportion cured.

4.4 Conclusions

In this chapter, two new distributions that can assume defective forms were proposed, the gamma-Gompertz and the gamma-inverse Gaussian. It is showed that can be induced new defective distributions when using the gamma frailty term. An approach for the regression version of these models and the inference by maximum likelihood approach were proposed. Simulation studies were presented to verify the asymptotic properties of the maximum likelihood estimator, which we conclude that the model does not present any relevant problems in terms of estimation. To illustrate the methodology, the proposed model was fitted to three real data sets. One of them, the breast cancer data, is a new data set that we are using for the first time. The proposed model was compared with the frailty mixture model. It was showed that the defective models induced by a frailty term is a competitive model for modeling the proportion of cured. Moreover, the defective model has advantage of it not being necessary to assume the existence of a proportion cured. And it is not necessary limit the range of the parameter in the estimation procedure.

Furthermore, when $a > 0$, we have the frailty models, gamma-Gompertz and gamma-inverse Gaussian. When $a < 0$, we have the defective gamma-Gompertz and the defective gamma-inverse Gaussian models, induced by the frailty term. We empathizes that we only have a frailty term in these models when $a > 0$, different than what is founded in the work of Kettunen *et al.* (1991).

CHAPTER 5

BETA-PG FAMILY

Seeking for more flexible distributions, many authors have been generalizing a baseline distribution using family of distributions. These families adds extra parameters in a baseline distribution in order to increase its modeling capability. For more information about this distribution families see Nadarajah, Rocha *et al.* (2015).

Using distribution families, in this chapter, is presented another approach to estimating the cure rate. Here a parameter "p" is added in the Beta-G family in order to create a new family of cure rate models. This family is a cure model, but it does not use the concept of defective distribution in its construction. This chapter is organized as follows. In Section 5.1, the Beta-G family of distribution is presented. In Section 5.2, the Beta-pG family of distribution is defined. In the next section, 5.3, an approach for the Beta-pG regression model is presented. In Section 5.5, the estimation issues using the maximum likelihood estimator is discussed. In the end, the methodology is finished by presenting a special case of the Beta-pG, the Beta-pExp, which is haved when G is taken as the exponential distribution.

5.1 The Beta-G Family of Distributions

The Beta distribution has been proved very useful in many contexts due to its flexibility. Its density function is defined by

$$f_{\beta}(t) = \frac{1}{B(a,b)} t^{a-1}(1-t)^{b-1},$$

where $B(a,b) = \int_0^1 v^{a-1}(1-v)^{b-1}dv$ denotes the Beta function, for $t \in (0,1)$, with $a > 0$ and $b > 0$ its shape parameters.

Eugene, Lee and Famoye (2002) proposed a general class of distributions generated from the Beta distribution. These distributions were motivated to model the failure time of a a-out-of-$a+b-1$ system when the failure times of the components are independent and identically distributed random variables.

The Beta-G family of distributions has its cumulative distribution function (cdf) defined as

$$F(t) = I_{G(t)}(a,b), \tag{5.1}$$

for $a > 0$, $b > 0$ and $G(\cdot)$ the cumulative distribution function of the baseline distribution, where

$$I_x(a,b) = \frac{\int_0^x v^{a-1}(1-v)^{b-1}dv}{B(a,b)},$$

denotes the incomplete Beta function ratio.

The probability density function (pdf) for these family is defined as

$$f(t) = \frac{1}{B(a,b)} g(t)\,[G(t)]^{a-1}\,[1-G(t)]^{b-1},$$

where $g(\cdot)$ is the density function of the baseline distribution. The corresponding survival and hazard functions are, respectively,

$$S(t) = 1 - I_{G(t)}(a,b) \quad \text{and} \quad h(t) = \frac{g(t)\,[G(t)]^{a-1}\,[1-G(t)]^{b-1}}{B(a,b)[1-I_{G(t)}(a,b)]}.$$

5.2 The Beta-pG Family of Distributions

The new family is presented by add a parameter p in the Beta-G family. In the mixture model, the parameter p only acts multiplicatively in the baseline distribution, in such way that the flexibility of the model totally relies in the flexibility of the baseline distribution. The parameter p only sets the "new zero" of the curve. Our proposal here is to compose this parameter in the Beta-G family in order to bring more flexibility than just multiply the baseline survival curve.

The purpose is to include $pG(t)$ in (5.1) instead to use just the function $G(t)$. Then, the cumulative density function of the Beta-pG family is defined as

$$F(t) = I_{pG(t)}(a,b) = \frac{\int_0^{pG(t)} v^{a-1}(1-v)^{b-1}dv}{B(a,b)},$$

where $a > 0$, $b > 0$, $p \in (0,1)$, $G(\cdot)$ is the cumulative distribution function of a baseline lifetime distribution and $g(\cdot)$ is its density function.

In this way, the probability density for this family is

$$f(t) = \frac{1}{B(a,b)} pg(t)\,[pG(t)]^{a-1}\,[1-pG(t)]^{b-1}.$$

The corresponding survival and hazard functions are, respectively,

$$S(t) = 1 - I_{pG(t)}(a,b) \quad \text{and} \quad h(t) = \frac{pg(t)\,[pG(t)]^{a-1}\,[1-pG(t)]^{b-1}}{B(a,b)[1-I_{pG(t)}(a,b)]}.$$

The proportion of the immune population is obtained by calculating the limit of survival function using the estimated parameters. Once that the cure rate depends on other parameters, the standard error is estimated by the delta method. The cure fraction p^* is easily calculated by the limit of the survival function

$$p^* = \lim_{t\to\infty} S(t) = \lim_{t\to\infty}[1 - I_{pG(t)}(a,b)] = 1 - I_p(a,b),$$

since $\lim_{t\to\infty} G(t) = 1$.

Some simulations studies and real data application will be used to show that the proposed family can lead to more adequate fits than the standard mixture model, using the same number of parameters, but just adding it in different ways. In the next section a regression approach for this family is proposed.

5.3 The Beta-pG Family of Regression Models

Here, it is briefly discussed the possibility of incorporating covariates into the Beta-pG model. For illustrative purposes, the parameter p is linked to explanatory variables through logit-link function, that is,

$$p(\boldsymbol{x}) = \frac{\exp\{\boldsymbol{x}^\top \boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}\}},$$

where $\boldsymbol{x}^\top = (1, x_1, \ldots, x_m)$ and $\boldsymbol{\beta}^\top = (\beta_1, \ldots, \beta_m)$ are the set of covariates and their regression coefficients.

Thus, the cumulative density function of the regression Beta-pG family is

$$F(t|\boldsymbol{x}) = I_{p(\boldsymbol{x})G(t)}(a,b).$$

The density function of the regression Beta-pG family is

$$f(t|\boldsymbol{x}) = \frac{1}{B(a,b)} p(\boldsymbol{x}) g(t) \left[p(\boldsymbol{x})G(t)\right]^{a-1} \left[1 - p(\boldsymbol{x})G(t)\right]^{b-1}.$$

The survival and hazard functions of the Beta-pG regression model are given by

$$S(t|\boldsymbol{x}) = 1 - I_{p(\boldsymbol{x})G(t)}(a,b) \quad \text{and} \quad h(t|\boldsymbol{x}) = \frac{p(\boldsymbol{x}) g(t) \left[p(\boldsymbol{x})G(t)\right]^{a-1} \left[1 - p(\boldsymbol{x})G(t)\right]^{b-1}}{B(a,b)[1 - I_{p(\boldsymbol{x})G(t)}(a,b)]}.$$

The cure fraction is calculated by

$$p^*(\boldsymbol{x}) = 1 - I_{p(\boldsymbol{x})}(a,b) = 1 - I_{\frac{\exp\{\boldsymbol{x}^\top \boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}^\top \boldsymbol{\beta}\}}}(a,b).$$

5.4 Especial cases

5.4.1 The Beta-pWeibull distribution

A particular case of the Beta-G family of distribution is the Beta-Weibull, which is obtained by taking $G(t)$ as the cdf of the Weibull distribution. This distribution was studied in detail by Lee, Famoye and Olumolade (2007) and the pdf, survival and hazard functions are, respectively,

$$\begin{aligned} f(t) &= \frac{\phi\lambda^{\phi}}{B(a,b)} t^{\phi-1} e^{-b(\lambda t)^{\phi}} \left[1-e^{-(\lambda t)^{\phi}}\right]^{a-1}, \\ S(t) &= 1 - I_{\left(1-e^{-(\lambda t)^{\phi}}\right)}(a,b), \end{aligned}$$

and

$$h(t) = \frac{\phi\lambda^{\phi} t^{\phi-1} e^{-b(\lambda t)^{\phi}} \left(1-e^{-\lambda t}\right)^{a-1}}{B(a,b)\left[1 - I_{\left(1-e^{-(\lambda t)^{\phi}}\right)}(a,b)\right]},$$

for $t > 0$, $a > 0$, $b > 0$, $\phi > 0$ and $\lambda > 0$.

Thus, the Beta-pWeibull distribution has density function given by

$$f(t) = \frac{\phi\lambda^{\phi}}{B(a,b)} t^{\phi-1} p e^{-(\lambda t)^{\phi}} \left[p\left(1-e^{-(\lambda t)^{\phi}}\right)\right]^{a-1} \left[1-p\left(1-e^{-(\lambda t)^{\phi}}\right)\right]^{b-1},$$

for $t > 0$, $a > 0$, $b > 0$, $\phi > 0$, $\lambda > 0$ and $p \in (0,1)$. The corresponding survival is given by

$$S(t) = 1 - I_{p\left(1-e^{-(\lambda t)^{\phi}}\right)}(a,b)$$

and the hazard function is

$$h(t) = \frac{\phi\lambda^{\phi} t^{\phi-1} p e^{-(\lambda t)^{\phi}} \left[p(1-e^{-(\lambda t)^{\phi}})\right]^{a-1} \left[1-p(1-e^{-(\lambda t)^{\phi}})\right]^{b-1}}{B(a,b)[1 - I_{p(1-e^{-(\lambda t)^{\phi}})}(a,b)]}.$$

Note that the Beta-pExponential (Beta-pExp) is a special case when $\phi = 1$. Figure 20 shows the pdf, survival and hazard functions of the Beta-pExp distribution for several parameters choice.

5.4.2 The Beta-pLindley distribution

Another special case of the Beta-G family of distribution is the Beta-Lindley, which is obtained by taking $G(t)$ as the cdf of the Lindley distribution. The Beta-Lindley was studied in

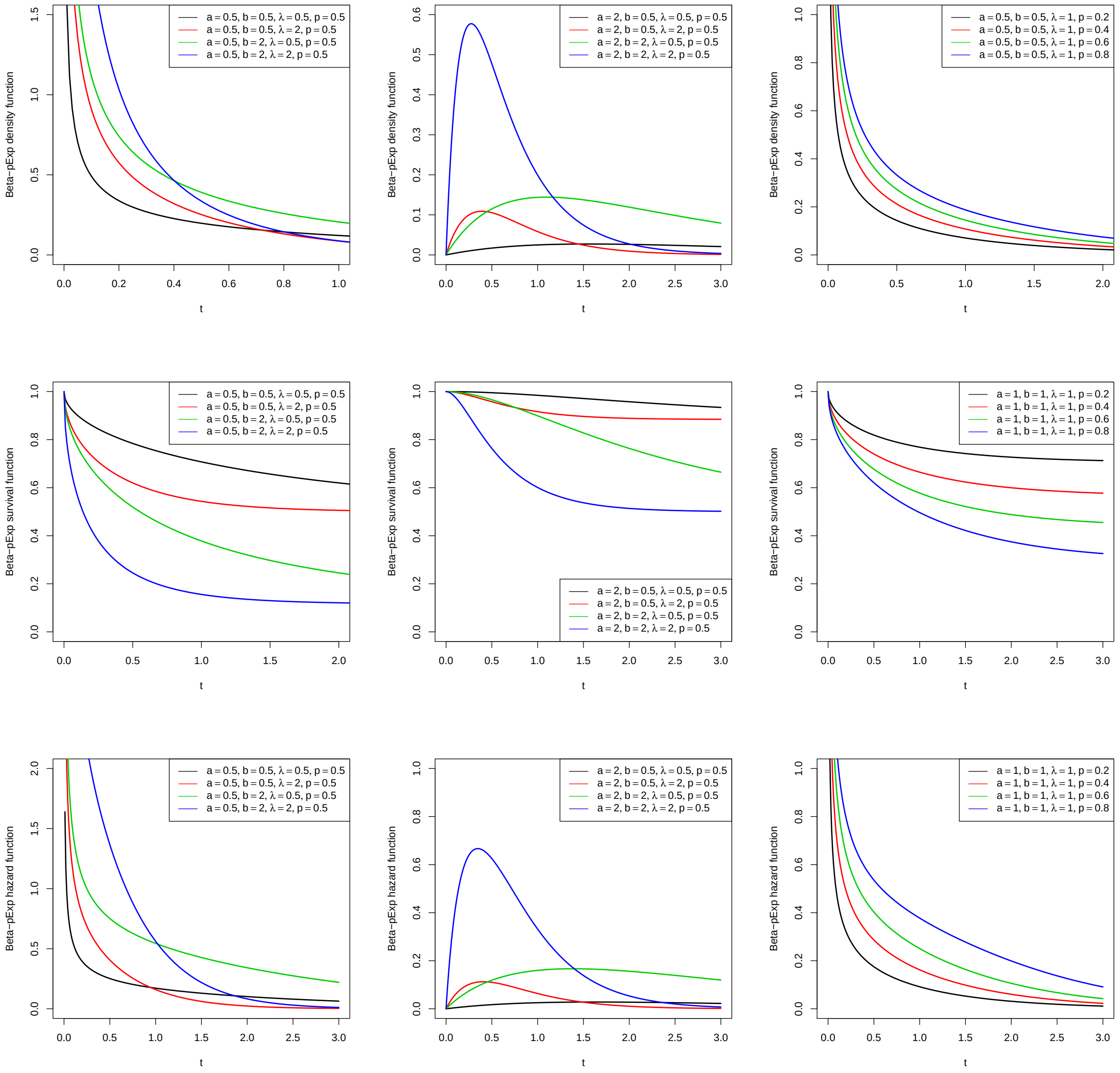

Figure 20 – Curves for the probability density (first line), survival function (second line) and hazard (third line) functions of the Beta-pExp distribution for several parameter choices.

detail by Merovci and Sharma (2014) and the pdf, survival and hazard functions are, respectively,

$$
\begin{aligned}
f(t) &= \frac{\lambda^2(\lambda+1+\lambda t)^{b-1}(1+t)e^{-\lambda bt}}{B(a,b)(\lambda+1)^b}\left(1-\frac{\lambda+1+\lambda t}{\lambda+1}e^{-\lambda t}\right)^{a-1},\\
S(t) &= 1-I_{\left(1-\frac{\lambda+1+\lambda t}{\lambda+1}e^{-\lambda t}\right)}(a,b)\\
h(t) &= \frac{\lambda^2(\lambda+1+\lambda t)^{b-1}(1+t)e^{-\lambda bt}}{B(a,b)(\lambda+1)^b\left(1-I_{\left(1-\frac{\lambda+1+\lambda t}{\lambda+1}e^{-\lambda t}\right)}(a,b)\right)}\left(1-\frac{\lambda+1+\lambda t}{\lambda+1}e^{-\lambda t}\right)^{a-1},
\end{aligned}
$$

for $t > 0$, $a > 0$, $b > 0$ and $\lambda > 0$.

Thus, the pdf, survival and hazard functions of the Beta-pLindley distribution are, respectively,

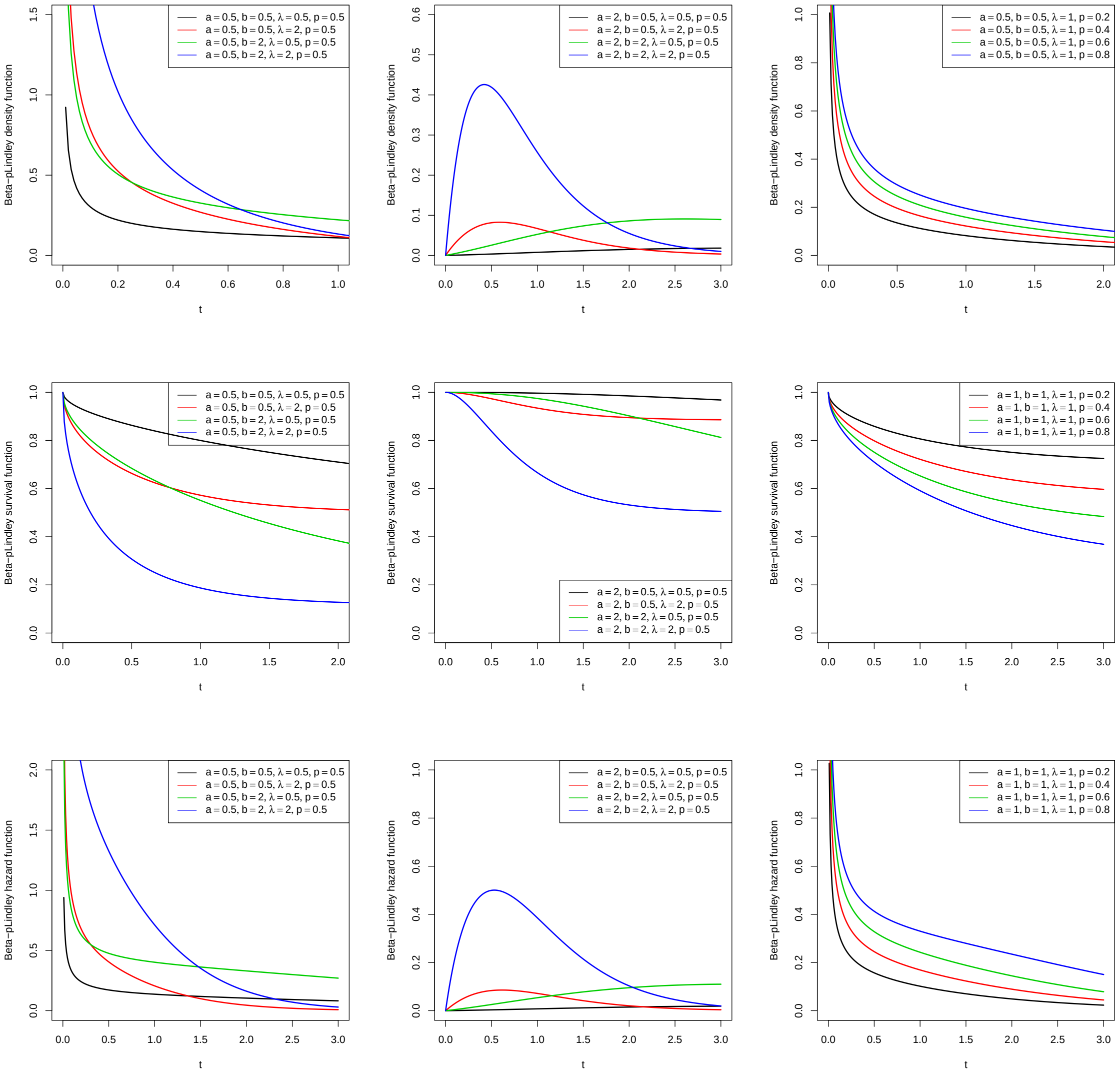

Figure 21 – Curves for the density (first line), survival (second line) and hazard (third line) functions of the Beta-pLindley distribution for several parameter choices.

$$
\begin{aligned}
f(t) &= \frac{p\lambda^2(1+t)e^{-\lambda t}}{B(a,b)(\lambda+1)}\left[p\left(1-\frac{\lambda+1+\lambda t}{\lambda+1}e^{-\lambda t}\right)\right]^{a-1}\left[1-p\left(1-\frac{\lambda+1+\lambda t}{\lambda+1}e^{-\lambda t}\right)\right]^{b-1},\\
S(t) &= 1-I_{p\left(1-\frac{\lambda+1+\lambda t}{\lambda+1}e^{-\lambda t}\right)}(a,b),\\
\text{and,}\\
h(t) &= \frac{p\lambda^2(1+t)e^{-\lambda t}}{B(a,b)(\lambda+1)\left[1-I_{p\left(1-\frac{\lambda+1+\lambda t}{\lambda+1}e^{-\lambda t}\right)}(a,b)\right]}\left[p\left(1-\frac{\lambda+1+\lambda t}{\lambda+1}e^{-\lambda t}\right)\right]^{a-1}\\
&\quad \left[1-p\left(1-\frac{\lambda+1+\lambda t}{\lambda+1}e^{-\lambda t}\right)\right]^{b-1}.
\end{aligned}
$$

for $t>0$, $a>0$, $b>0$, $\lambda>0$ and $p\in(0,1)$.

5.4.3 The Beta-pGamma distribution

The Beta-pGamma distribution is another case of the Beta-G family, which is obtained by taking $G(t)$ as the cdf of the Gamma distribution. The Beta-Gamma distribution was studied in detail by Kong, Carl and Sepanski (2007) and the pdf, survival and hazard functions are, respectively, given by

$$f(t) = \frac{\phi^{\lambda} t^{\lambda-1} e^{-\phi t}}{B(a,b)\Gamma(\lambda)} \left[G(t)\right]^{a-1} \left[1-G(t)\right]^{b-1}, \quad S(t) = 1 - I_{G(t)}(a,b),$$

and,

$$h(t) = \frac{\phi^{\lambda} t^{\lambda-1} e^{-\phi t}}{B(a,b)\Gamma(\lambda)\left[1 - I_{G(t)}(a,b)\right]} \left[G(t)\right]^{a-1} \left[1-G(t)\right]^{b-1},$$

where $t > 0$, $a > 0$, $b > 0$, $\lambda > 0$, $\phi > 0$ and $G(\cdot)$ is the cdf of the Gamma distribution with parameters λ and ϕ.

This way, the pdf, survival and hazard functions of Beta-pGamma distribution are, respectively,

$$f(t) = \frac{p\phi^{\lambda} t^{\lambda-1} e^{-\phi t}}{B(a,b)\Gamma(\lambda)} \left[pG(t)\right]^{a-1} \left[1-pG(t)\right]^{b-1}, \quad S(t) = 1 - I_{pG(t)}(a,b),$$

and

$$h(t) = \frac{p\phi^{\lambda} t^{\lambda-1} e^{-\phi t}}{B(a,b)\Gamma(\lambda)\left[1 - I_{(pG(t))}(a,b)\right]} \left[pG(t)\right]^{a-1} \left[1-pG(t)\right]^{b-1}.$$

Due the several possibilities of choices of the G distribution, the Beta-pG family of distributions becomes flexible to accommodate several kinds of survival data with a proportion surviving.

5.5 Inference

In this section, the inferential procedure is described, which is based on the maximum likelihood approach and the asymptotic large sample theory. Consider T_i a random variable that denotes the time of the event of interest and C_i the censoring time for the ith individual. For the ith observation the observed time is $t_i = \min\{T_i, C_i\}$ with $\delta_i = \mathbb{I}(T_i \leq C_i)$, where $\delta_i = 1$ if t_i is a failure time and $\delta_i = 0$ if it is right censored, for $i = 1, \ldots, n$. The observed data is represented by $D = (t, \boldsymbol{\delta}, \mathbf{X})$, where $t = (t_1, \ldots, t_n)^\top$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$ and $\mathrm{X} = (x_1, \ldots, x_m)^\top$ is an $n \times m$ matrix containing the covariates. Suppose that the data is independently and identically distributed and come from a distribution with probability density and survival functions specified by $f(\cdot; \boldsymbol{\theta})$ and $S(\cdot; \boldsymbol{\theta})$, respectively, where $\boldsymbol{\theta} = (a, b, p, \boldsymbol{\gamma})^\top$ denotes a vector of parameters and $\boldsymbol{\gamma}$ denotes the vector of parameters of the baseline distribution. The density and cumulative function of the baseline distribution are given by $g(t; \boldsymbol{\gamma})$ and $G(t; \boldsymbol{\gamma})$, respectively.

The likelihood function of parameters $\boldsymbol{\theta}$ under non informative censoring can be written as klein2003,

$$L(\boldsymbol{\theta};D) \propto \prod_{i=1}^{n} f(t_i;\boldsymbol{\theta})^{\delta_i} S(t_i;\boldsymbol{\theta})^{1-\delta_i}.$$

The corresponding log-likelihood function $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta};D)$ is

$$\ell(\boldsymbol{\theta}) = const + \sum_{i=1}^{n} \delta_i \log f(t_i;\boldsymbol{\theta}) + \sum_{i=1}^{n} (1-\delta_i) \log S(t_i;\boldsymbol{\theta}).$$

Thus, for the Beta-pG regression model the likelihood function for $\boldsymbol{\theta}$ is

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= const + \sum_{i=1}^{n} \delta_i \left[\log\left(\frac{1}{B(a,b)} p(\boldsymbol{x}) g(t_i;\boldsymbol{\gamma}) \left[p(\boldsymbol{x})G(t_i;\boldsymbol{\gamma})\right]^{a-1} \left[1-p(\boldsymbol{x})G(t_i;\boldsymbol{\gamma})\right]^{b-1} \right) \right] \\ &+ \sum_{i=1}^{n} (1-\delta_i) \left[\log\left(1 - I_{p(\boldsymbol{x})G(t_i;\boldsymbol{\gamma})}(a,b)\right)\right]. \end{aligned} \tag{5.2}$$

The maximum likelihood estimator is the solution of the system

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

where $\boldsymbol{\theta} = (a,b,p,\boldsymbol{\gamma})^\top$.

The maximum likelihood estimates (MLEs) of the parameters are obtained by numerically maximizing the log-likelihood function (5.2). There are many methods available for numerical maximization. The routine optim in the R software for numerical maximization was used (R Core Team, 2013).

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

$$\widehat{\Sigma}(\widehat{\theta}) = \left\{ -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^\top}\bigg|_{\theta=\widehat{\theta}} \right\}^{-1}.$$

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

In the next section, an extensive simulation studies are performed to investigate numerically the asymptotic properties of the MLEs, as well as the performance of the proposed models in terms of the AIC in comparison with their own version as a standard mixture model.

5.6 Simulation Studies

Two simulation experiments were performed. The first one is to assess the performance of the MLEs of the proposed model parameters, considering different sample sizes. In the second one, the Beta-pExp and the mixture Beta-Exp distributions are compared in terms of the AIC criterion and cure fraction point estimation, when the data is generated from the Beta-pExp distribution. All computations were performed in R (R Core Team, 2013). A random sample of size n containing true times, censored times and a binary covariate X were generated, which leads to the cured fraction p_0 and p_1 for the two levels of X, respectively. For that, consider the algorithm, describe in Section 2.8, to generate the data.

In the first experiment, the average of mean square error of the MLEs, the bias, the coverage probability of 95% confidence intervals and coverage length for each parameter were calculated. Different sample sizes as $n = 100, 300, \ldots, 2500$ were considered. In each scenario (each combination of parameters values and sample size), 200 random samples were performed.The scenarios are defined by the following parameter choices $(a, b, \lambda, \beta_1, p_0, p_1)$:

- Scenario one: $(1, 0.5, 1, 1, 0.7071, 0.5185)$;

- Scenario two: $(1.5, 1, 1, 1, 0.6464, 0.0.3749)$;

- Scenario three: $(1.5, 1.5, 1, 1, 0.500, 0.2166)$.

These scenarios show a variety in cure rates values. Figure 22 shows the plots of the mean square errors, biases, coverage probabilities and coverage lengths of the parameters versus n for simulated data from proposed model. The red, green and blue lines (square, circle and triangle) represent the scenarios one, two and three, respectively.

It is worth mentioning that, on average, the mean square errors decreases fast as the sample size increases and gets reasonably close to zero for $n > 500$. The biases stays around zero for all parameters and the empirical coverage probabilities for all parameters are reasonably close to the nominal level. In all scenarios the coverage lengths decreases as the sample size increases, as expected.

For comparison of the models, the difference between AICs and BICs of fitted Beta-pExp mixture models and Beta-pExp was considered. This way, a positive AIC mean difference means that, on average, the AIC of the fitted Beta-pExp model is smaller than the AIC obtained from the fitted mixture Beta-pExp model, which shows the advantage of the proposed model.

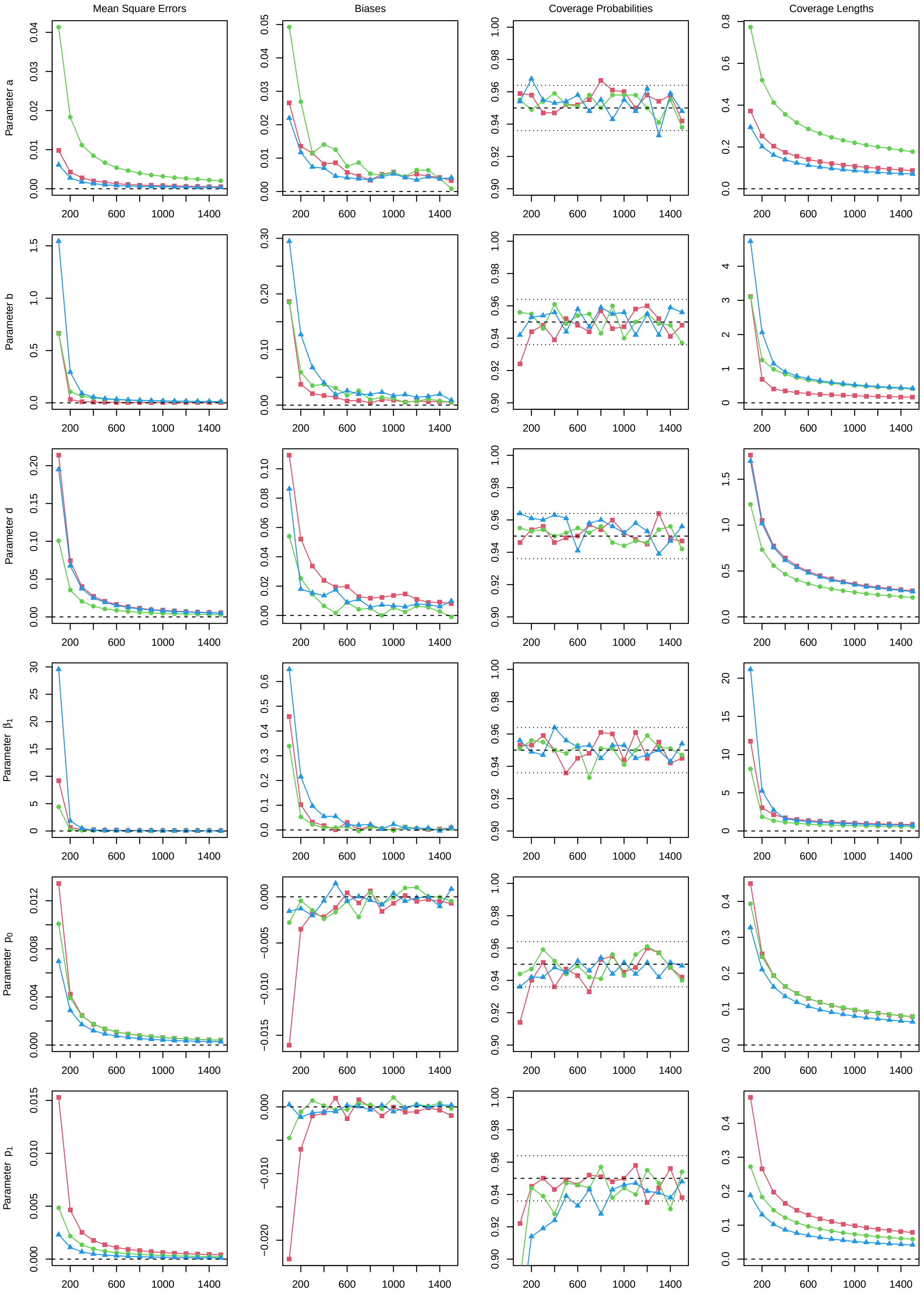

Figure 22 – Simulation of the Beta-pExp distribution. The red, green and blue (square, circle and triangle) represent scenarios one, two and three, respectively.

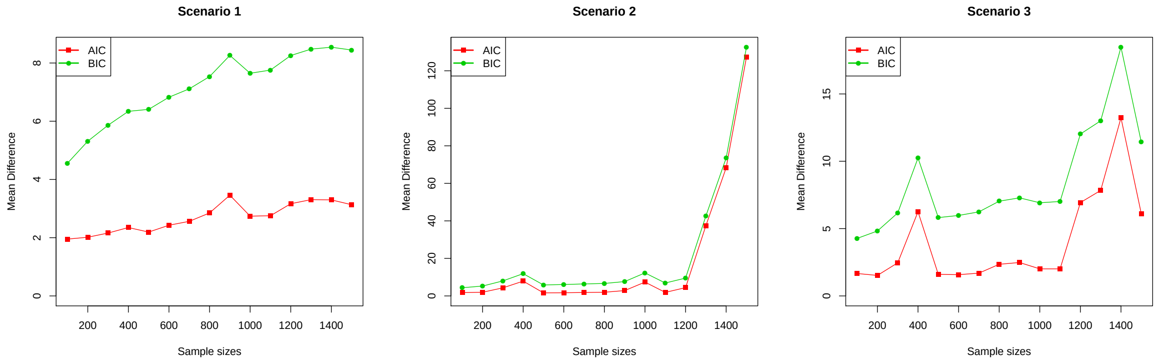

Figure 23 – AIC and BIC mean difference between the mixture Beta-Exp and Beta-pExp distributions. The left, mid and right panels represents the first, second and third scenarios, respectively.

The second experiment compares in term of AIC and BIC measure the proposed model with their corresponding mixture model. The data sets are generated from the Beta-pExp distribution. For a fixed scenario, the mean difference between AICs and BICs obtained of fitted Beta-pExp and mixture Beta-Exp models was evaluated. Was considered $n = 100, 200, \ldots, 1500$ and performed 200 simulations for each configuration. The scenarios are defined by the following parameter choices $(a, b, \lambda, \beta_1, p_0, p_1)$:

- Scenario one: $(1, 0.5, 1, 1, 0.7071, 0.5186)$;
- Scenario two: $(0.5, 1, 1, 1, 0.6464, 0.3749)$;
- Scenario three: $(0.5, 0.5, 1, 1, 0.500, 0.2166)$.

Figure 23 shows the difference between the AIC and BIC values of the mixture Beta-Exp and Beta-pExp distributions. This way, if the mean difference is positive, then the Beta-pExp distributions have a smaller AIC and BIC criteria compared to the mixture Beta-Exp distribution. It was noticed that, in all scenarios, the AIC and BIC mean difference increases with the sample size. It starts around 1.5 and finishes with a difference greater than hundred.

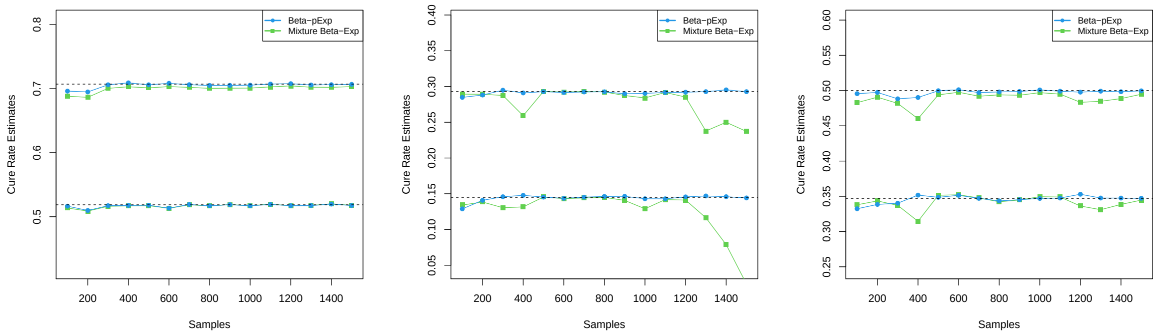

Figure 24 – Cure rate estimates for each group obtained from both models. The left, mid and right panels represents the first, second and third scenarios, respectively.

The point estimates of the cure rates were also compared. Figure 24 shows the estimates of the cured fraction p_0 and p_1 from the Beta-pExp (blue line with squares) and mixture Beta-Exp (green line with circles) distributions. The results show that the point estimate is quite precise for both distributions, but in some cases, the Beta-pExp is clearly more accurate and is least affected with the changed of the samples size. For example, in the estimation of $p_0 = 0.519$ in the first scenario and $p_0 = 0.347$ in the third scenario.

5.7 Applications

To illustrate the proposed model, two real cancer datasets were considered. The Beta-pG and mixture Beta-p models were fitted and compared with survival curve estimates obtained using the Kaplan-Meier estimator. For each fitted model, the maximum likelihood estimator, standard error, 95% confidence interval estimates for the parameters and AIC, BIC, and CAIC values were provided. The delta method was used to estimate the standard error for cure rate parameter.

The mixture model is obtained by taking the Beta-Exp distribution as $S_0(t)$ in (1.1). The Beta-Exp mixture model is defined as

$$S(t|\boldsymbol{x}) = p(x) + (1 - p(x)) \left(1 - I_{\left(1-e^{-(\lambda t)^{\phi}}\right)}(a,b)\right)$$

5.7.1 Melanoma data

This data set arises from a melanoma studies from the Eastern Cooperative Oncology Group. They tested the efficacy of a drug administered for one year to prevent relapse and death of high risk patients after a curative surgery for melanoma. The data set has 285 patients, of which 98 (34.39%) are censored. The control group has 140 (49.12%) observations, while the group under treatment has 145 patients. Besides of the observed time and censoring indicator, other variables were measured at baseline, such as sex, age and treatment. For illustrative purposes, only the treatment (TRT) is considered as covariate. The control group is represented by 0 and the group under treatment is denoted by 1. For more information on this data set, please check Kirkwood *et al.* (1996).

Table 12 – Maximum likelihood estimates. standard error (SE). 95% confidence interval and AIC values obtained by Beta-pExp and Mixture Beta-Exp models fits for the melanoma data set.

	Beta-pExp model				Mixture Beta-Exp Model			
Parameter	Estimate	SE	Lower CI	Upper CI	Estimate	SE	Lower CI	Upper CI
a	1.062	0.1164	0.8339	1.2901	0.9218	0.0847	0.7557	1.0879
b	2.2769	0.3116	1.6662	2.8876	3.2579	10.3098	-16.949	23.4647
λ	0.764	0.0839	0.5995	0.9285	0.2506	0.8032	-1.3235	1.8248
β_0	-	-	-	-	-1.1417	0.2016	-1.5368	-0.7466
β_1	-0.48	0.1816	-0.8359	-0.1242	0.5409	0.2679	0.0159	1.066
p_0	0.2213	0.0349	0.1529	0.2896	0.242	0.037	0.1695	0.3145
p_1	0.3549	0.0402	0.2761	0.4336	0.3542	0.0405	0.2747	0.4336
AIC	772.52				780.88			
BIC	801.22				816.75			
CAIC	805.22				821.75			

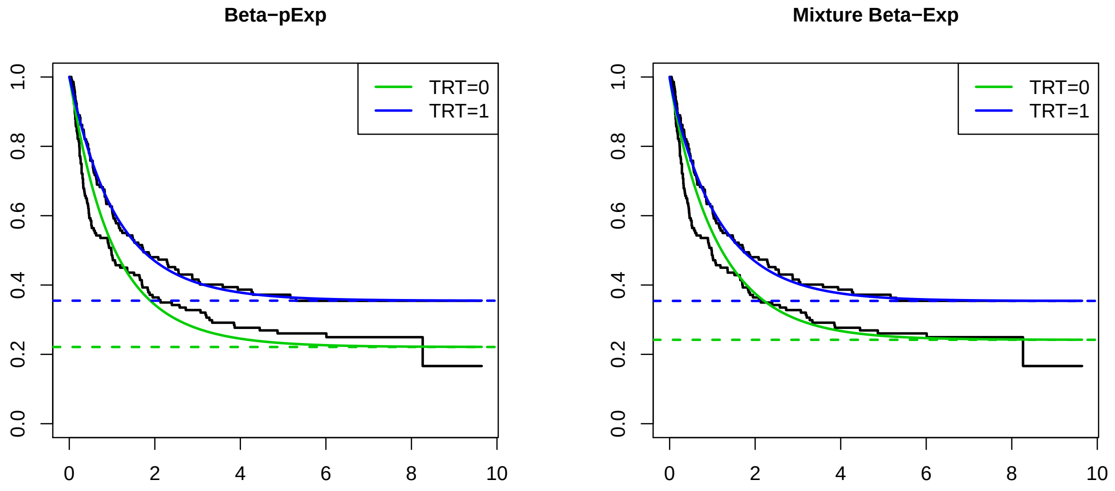

Figure 25 – Kaplan–Meier survival estimates curves (black line) stratified by treatment (0: control group; 1: treatment group) and survival function estimates according to different models.

The fit of the Beta-pExp and the mixture Beta-Exp models are compared using the AIC, BIC and CAIC criterion measures. As a visual assessment, the plot of the fitted models together with the Kaplan-Meier non-parametric estimator curve were provided (KAPLAN; MEIER, 1958).

Table 12 shows the MLEs of the Beta-pExp and mixture Beta-Exp models for the melanoma data set. Survival functions estimates are presented in Figure 25. Notice that both models present quite close estimates for p_1, around 0.35, whereas for the parameter p_0 there is a slight difference. According to the AIC, BIC and CAIC criterion, the Beta-pExp model seems to be the best choice. Additionally, both models indicate a significant effect of treatment (zero is not included in the 95% confidence interval of parameter β_1). However, there is not a significant difference between the cure rates for the group under treatment and the control group. In the

Beta-pExp, $\hat{p}_0 = 0.22$ and in the mixture Beta-Exp, $\hat{p}_0 = 0.24$.

Figure 25 presents the fitted survival curves in the data. Not as clearly, but the same happens for $TRT = 0$x'. Therefore, it can be concluded that the Beta-pExp model is more appropriate for this data than the mixture Beta-Exp. Observe that the survival time of the treatment group ($TRT = 1$) is higher than the control group ($TRT = 0$), and differs by around 7%. And the results suggest a effect of treatment in lifetime, regardless of model, observe that the interval the β_1 does not include 0. And cure rate is estimated in $p_0 = 0.22$ with standard error 0.03 ($TRT = 0$) and $p_1 = 0.35$ with standard error 0.04 ($TRT = 1$).

The closer the parametric models gets to the Kaplan-Meier curve, the better the fit is. In both models the delta method was used to calculate the standard error of the cure rate.

5.7.2 Oncocentro cancer data

This data set was provided by the Fundação Oncocentro de São Paulo (FOSP), which is responsible for coordinating the Hospital Cancer Registry of the State of São Paulo. The FOSP is a public institution connected to the State Health Secretariat, which assists in the preparation and implementation of healthcare policies in the field of Oncology, and serves as an instrument so that oncology hospitals can prepare their own protocols and improve their care practices (for more information see Calsavara *et al.* (2019b)). This data contain information about the failure time or censoring (in years) from 7166 patients diagnosed with melanoma in the state of São Paulo, Brazil, between 2000 and 2014, with follow-up conducted until 2018. The event of interest is death by melanoma cancer and the data set has 5099 patients are censored (71.15%). Besides of the observed time and censoring indicator, other variables were measured at baseline, such as sex, age, surgery and treatment. For illustrative purposes, only the sex is considered as covariate.

Table 13 – Maximum likelihood estimates. standard error (SE). 95% confidence interval and AIC. BIC. CAIC values obtained by Beta-pExp and Mixture Beta-Exp models fits for the oncocentro data set.

	Beta-pExp Model				Mixture Beta-Exp Model			
	Estimate	SE	Lower IC	Upper IC	Estimate	SE	Lower IC	Upper IC
a	1.007	0.032	0.945	1.069	0.988	0.045	0.900	1.075
b	1.522	0.107	1.311	1.732	0.744	2.576	-4.305	5.793
	0.263	0.015	0.234	0.293	0.374	1.316	-2.205	2.953
β_0	-	-	-	-	-0.545	0.105	-0.751	-0.339
β_1	-0.586	0.049	-0.681	-0.491	0.620	0.062	0.499	0.741
p_0	0.666	0.011	0.645	0.687	0.667	0.011	0.646	0.689
p_1	0.513	0.013	0.487	0.538	0.519	0.014	0.492	0.545
AIC	14885.269					14894.064		
BIC	14912.777					14928.450		
CAIC	14873.271					22321.096		

The main goal was to assess the impact of sex on specific survival. Of the 7166 patients, 3538 patients are male (49.28%) and 3634 patients are female (50.71%). A total of 2067 patients death by cancer (28.84%), events occurred during follow-up period: 1188 occurred among male patients (represents 50.68% of male patients) and 879 female patients (represents 31.90% of female patients). The maximum observation time was approximately 18.54 years.

The results of the fitted Beta-pExp e Mixture Beta-Exp models are showed in Table 13. According to the AIC, BIC and CAIC values, the Beta-pExp model seems to be the better choice among the models. The results suggest that the sex influences in lifetime and the cure rate estimated in the models are similar. The cure rate estimated are $p_0 = 0.66$ with standard error 0.01 (female) and $p_1 = 0.51$ with standard error 0.013 (male). Overall, the models reasonably fit Kaplan–Meier curves.

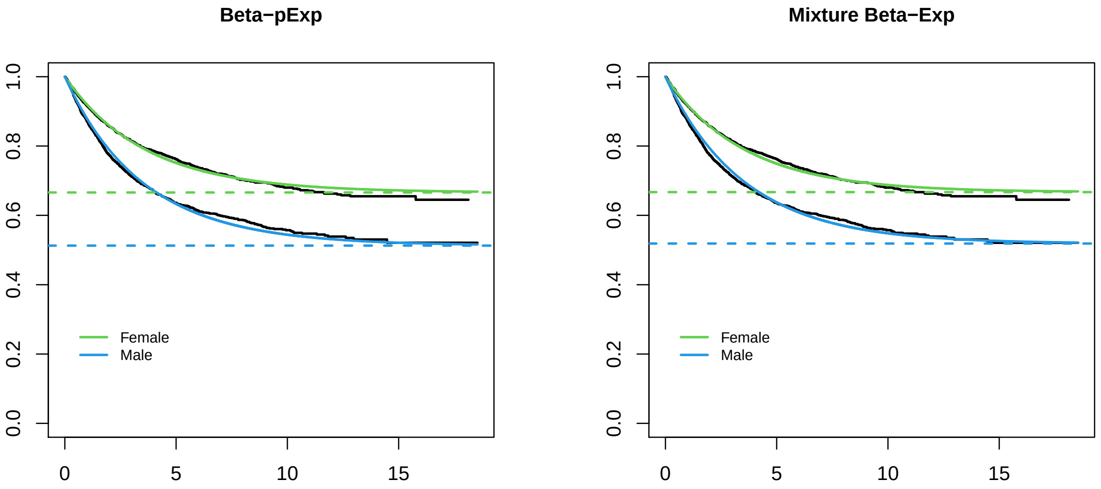

Figure 26 – Kaplan–Meier survival estimates curves (black line) stratified by sex covariate and survival function estimates according to different models.

5.8 Conclusions

In this chapter, a new family of cure rate models was proposed, the Beta-pG family of distributions. A regression model was proposed to accommodate covariate information in the family. Some special case the Beta-pG family were considered as, Beta-pWeibull, Beta-pLindley and Beta-pGamma. It was considered the special case when G comes from an exponential distribution for the simulation studies. Simulation study to illustrate frequentist properties of the maximum likelihood estimators of the proposed model parameters, where the mean squared error appears reasonably close to 0 as sample size increases, for all estimators. As in practice situations the choice of model is often based on a selection criterion, the performance of the model in terms of AIC, BIC and CAIC measures was evaluated to compare with the mixture model. It was observed that, in average, the proposed model outperfoms the respective mixture model in terms of criteria measure, when the data is generated from the Beta-pExp distribution. This new family was showed to be an alternative model to calculate the cure rate. The practical relevance and applicability of the proposed model is illustrated in a real data set, in which our model yields a slight better fit than the mixture model. This generalization is expected to attract wider applications in survival analysis.

CHAPTER

6

FINAL REMARKS

6.1 Conclusions

An important area of survival analysis is related to cure rate models (or long-term survival models). Cure rate models basically focus on the proportion of patients who survive long-term following disease. Additionally, these models focus on the probability of survival of the uncured patients up to a given point in time. Several extensions of these models have also been investigated. So the main contribution of this thesis was to approach two alternative methods to model the cure rate in long-term models.

The first use the methodology of the defective models that has the advantage to modeling the proportion of cured without adding any extra parameters in the model, in contrast to the most models from the literature. In Chapter 3, the Gompertz and inverse Gaussian defective models were presented. Simulation studies were used to check the performance of the maximum likelihood estimators for these models. In Chapter 4, two new distributions were proposed, the defective Gamma-Gompertz and the defective Gaussian-inverse gamma. It was show that its possible to induce new defective distributions when using the gamma frailty term. In addition, a version to include covariables in these models and their estimation process are presented. The properties of these models through simulation studies were analyzed and an application was presented to illustrate the proposed models. The models presented showed to be as efficient as the standard mixture model to model the survival data with cure rate.

The second method proposed use the concept the distributions family. These families adds extra parameters in a baseline distribution in order to increase its modeling capability. This work is proposed to include a parameter ’p’ in the Beta-G family (Chapter 5. Then, the Beta-pG cure rate models are generated. Some special cases of this family were presented and, in addition, an approach to introduce covariates in these models was discussed. Simulation studies have shown that there are no major problems in terms of estimates and that the proposed model is an alternative method to estimate the cure rate. Real data were used to illustrate the proposed model and to compare it with the standard mixture model. It was concluded that this new family of cure rate models is as flexible as the mixture model for estimating the cure fraction.

Summarizing, this thesis is a work related to the cure rate modeling in survival studies using different methods to estimate the cure rate. In Chapter 4, two new defective distributions were shown. And in Chapter 5, a new family distribution was proposed to estimate the cure rate, and three new cure rate models were presented, but more cure rate models can be generated using this method.

This thesis is based on two papers previously developed works of the author. One is already published (SCUDILIO *et al.*, 2019) and the other is submitted.

6.2 Future Works

Here are presented suggestions for posterior works:

- To propose a new defective family distribution using two results presented by Rocha (2016).
- To develop another defective distribution induced by a different fragility distribution;
- To propose new families of long-term models using another family of distributions, for example, the Gama-G family;
- To present simulation studies from the Bayesian point of view for the models proposed.

6.3 Acknowledgments

Thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001 for financial support during the course of this PhD.

BIBLIOGRAPHY

AALEN, O. O. Modelling heterogeneity in survival analysis by the compound Poisson distribution. **Annals of Applied Probability**, v. 4, p. 951–972, 1992. Citations on pages 27 and 49.

AALEN, O. O.; GJESSING, H. K. Understanding the shape of the hazard rate: A process point of view (with comments and a rejoinder by the authors). **Statistical Science**, Institute of Mathematical Statistics, v. 16, n. 1, p. 1–22, 2001. Citation on page 21.

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, IEEE, v. 19, n. 6, p. 716–723, 1974. Citation on page 30.

ALEXANDER, C.; CORDEIRO, G.; ORTEGA, E.; SARABIA, J. Generalized Beta-Generated Distributions. **Computational Statistics and Data Analysis**, v. 56, p. 1880–1897, 2012. Citation on page 21.

ALZAATREH, A.; LEE, C.; FAMOYE, F. A New Method for Generating Families of Continuous Distributions. **Metron**, v. 71, p. 63–79, 2013. Citation on page 21.

ANDERSEN, P.; BORGAN, O.; GILL, R.; KEIDING, N. Statistical models based on counting processes. **NY Springer**, 1993. Citation on page 21.

BALKA, J.; DESMOND, A. F.; MCNICHOLAS, P. D. Review and implementation of cure models based on first hitting times for wiener processes. **Lifetime Data Analysis**, Springer Verlag, New York, v. 15, n. 2, p. 147–176, 2009. Citations on pages 19 and 20.

____. Bayesian and likelihood inference for cure rates based on defective inverse gaussian regression models. **Journal of Applied Statistics**, Taylor and Francis, v. 38, n. 1, p. 127–144, 2011. Citation on page 20.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, Taylor and Francis, v. 47, n. 259, p. 501–515, 1952. Citation on page 19.

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 19.

BORGES, P. Em algorithm-based likelihood estimation for a generalized gompertz regression model in presence of survival data with long-term survivors: an application to uterine cervical cancer data. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 87, n. 9, p. 1712–1722, 2017. Citation on page 20.

CALSAVARA, V. F.; MILANI, E. A.; BERTOLLI, E.; TOMAZELLA, V. Long-term frailty modeling using a non-proportional hazards model: Application with a melanoma dataset. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, p. 0962280219883905, 2019. Citation on page 21.

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, 2019. Citations on pages 20 and 82.

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, p. 1–26, 2019. Citation on page 20.

CALSAVARA, V. F.; RODRIGUES, A. S.; TOMAZELLA, V. L. D.; CASTRO, M. de. Frailty models power variance function with cure fraction and latent risk factors negative binomial. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 46, n. 19, p. 9763–9776, 2017. Citation on page 21.

CANTOR, A. B.; SHUSTER, J. J. Parametric versus non-parametric methods for estimating cure rates based on censored survival data. **Statistics in Medicine**, Wiley Online Library, v. 11, n. 7, p. 931–937, 1992. Citation on page 20.

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 and Francis, v. 94, n. 447, p. 909–919, 1999. Citation on page 19.

CLAYTON, D. G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. **Biometrika**, Biometrika Trust, v. 65, n. 1, p. 141–151, 1978. Citations on pages 21, 27, and 49.

CORDEIRO, G.; CASTRO, M. A New Family of Generalized Distributions. **Journal of Statistical Computation and Simulation**, v. 81, p. 883–898, 2011. Citation on page 21.

CORDEIRO, G.; ORTEGA, E.; CUNHA, D. da. The Exponentiated Generalized Class of Distributions. **Journal of Data Science**, v. 11, p. 1–27, 2013. Citation on page 21.

CORDEIRO, G.; ORTEGA, E.; SILVA, G. The Beta Extended Weibull Family. **Journal of Probability and Statistical Science**, v. 10, p. 15–40, 2012. Citation on page 21.

COX, D. R. Regression models and life-tables (with discussion). **Journal of the Royal Statistical Society**, B, n. 34(2), 1972. Citation on page 27.

ELBERS, C.; RIDDER, G. True and spurious duration dependence: The identifiability of the proportional hazard model. **The Review of Economic Studies**, Oxford University Press, v. 49, n. 3, p. 403–409, 1982. Citations on pages 28 and 50.

EUGENE, N.; LEE, C.; FAMOYE, F. Beta-Normal Distribution and Its Applications. **Communications in Statistics—Theory and Methods**, v. 31, p. 497–512, 2002. Citations on pages 21 and 69.

FELLER, W. **An Introduction to Probability Theory, volumes I and II**. [S.l.]: John Wiley and Sons, New York, 1968. Citation on page 26.

GIESER, P. W.; CHANG, M. N.; RAO, P.; SHUSTER, J. J.; PULLEN, J. Modelling cure rates using the gompertz model with covariate information. **Statistics in medicine**, Wiley Online Library, v. 17, n. 8, p. 831–839, 1998. Citations on pages 20 and 34.

GUPTA, R.; GUPTA, P.; GUPTA, R. Modeling Failure Time Data by Lehman Alternatives. **Communications in Statistics—Theory and Methods**, v. 27, p. 887–904, 1998. Citation on page 21.

HAYBITTLE, J. L. The estimation of the proportion of patients cured after treatment for cancer of the breast. **British Journal of Radiology**, British Institute of Radiology, v. 32, n. 383, p. 725–733, 1959. Citation on page 20.

HOUGAARD, P. . . Modelling heterogeneity in survival data. **Journal of Applied Probability**, JSTOR, p. 695–701, 1991. Citation on page 20.

____. Frailty models for survival data. **Lifetime data analysis**, Springer, v. 1, n. 3, p. 255–273, 1995. Citation on page 21.

HOUGAARD, P. . Survival models for heterogeneous populations derived from stable distributions. **Biometrika**, v. 73, p. 387–396, 1986. Citations on pages 27 and 49.

HOUGAARD, P. A class of multivanate failure time distributions. **Biometrika**, v. 73, p. 671–678, 1986. Citations on pages 27 and 49.

HOUGAARD, P.; MYGLEGAARD, P.; BORCH-JOHNSEN, K. Heterogeneity models of disease susceptibility, with application to diabetic nephropathy. **Biometrics**, JSTOR, p. 1178–1188, 1994. Citation on page 21.

IBRAHIM, J. G.; CHEN, M.-H.; SINHA, D. Bayesian semiparametric models for survival data with a cure fraction. **Biometrics**, Wiley Online Library, v. 57, n. 2, p. 383–388, 2001. Citation on page 47.

JR, I. M. L.; HALLORAN, M. E. A frailty mixture model for estimating vaccine efficacy. **Applied Statistics**, JSTOR, p. 165–173, 1996. Citation on page 21.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, Taylor and Francis, v. 53, n. 282, p. 457–481, 1958. Citations on pages 25 and 81.

KETTUNEN, J. *et al.* **Transition Intensities from Unemployment**. [S.l.], 1991. Citations on pages 51 and 67.

KIRKWOOD, J. M.; STRAWDERMAN, M. H.; ERNSTOFF, M. S.; SMITH, T. J.; BORDEN, E. C.; BLUM, R. H. Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the eastern cooperative oncology group trial est 1684. **Journal of clinical oncology**, American Society of Clinical Oncology, v. 14, n. 1, p. 7–17, 1996. Citation on page 80.

KONG, L.; CARL, L.; SEPANSKI, J. On the Properties of Beta Gamma Distribution. **Journal of Modern Applied Statistical Methods**, v. 6, 2007. Citation on page 75.

LAURIE, J. A.; MOERTEL, C. G.; FLEMING, T. R.; WIEAND, H. S.; LEIGH, J. E.; RUBIN, J.; MCCORMACK, G. W.; GERSTNER, J. B.; KROOK, J. E.; MALLIARD, J. Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil. the north central cancer treatment group and the mayo clinic. **Journal of Clinical Oncology**, American Society of Clinical Oncology, v. 7, n. 10, p. 1447–1456, 1989. Citations on pages 20, 42, and 61.

LEE, C.; FAMOYE, F.; OLUMOLADE, O. Beta-weibull distribution: some properties and applications to censored data. **Journal of modern applied statistical methods**, v. 6, n. 1, p. 17, 2007. Citation on page 72.

LILLARD, L. A.; PANIS, C. W. aml multilevel multiprocess statistical software, release 1.0. **Los Angeles: EconWare**, 2000. Citation on page 43.

MARTINEZ, E. Z.; ACHCAR, J. A. The defective generalized gompertz distribution and its use in the analysis of lifetime data in presence of cure fraction, censored data and covariates. **Electronic Journal of Applied Statistical Analysis**, v. 10, n. 2, p. 463–484, 2017. Citations on pages 20 and 48.

____. A new straightforward defective distribution for survival analysis in the presence of a cure fraction. **Journal of Statistical Theory and Practice**, Taylor & Francis, v. 12, n. 4, p. 688–703, 2018. Citations on pages 20 and 48.

MEROVCI, F.; SHARMA, V. K. The beta-lindley distribution: properties and applications. **Journal of Applied Mathematics**, Hindawi Publishing Corporation, v. 2014, 2014. Citation on page 73.

MISSOV, T. I. Analytic expressions for life expectancy in gamma-gompertz mortality settings. 2010. Citations on pages 27 and 49.

____. Gamma-gompertz life expectancy at birth. **Demographic research**, v. 28, p. 259–270, 2013. Citations on pages 27 and 49.

NADARAJAH, S.; ROCHA, R. *et al.* Newdistns: An r package for new families of distributions. **Jnl Stat Soft. To appear.[Links]**, 2015. Citation on page 69.

OAKES, D. A model for association in bivariate survival data. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 414–422, 1982. Citation on page 21.

OEHLERT, G. W. A note on the delta method. **The American Statistician**, Taylor & Francis, v. 46, n. 1, p. 27–29, 1992. Citation on page 30.

PAPANICOLAOU, A. Taylor approximation and the delta method. **April**, v. 28, p. 2009, 2009. Citation on page 30.

PARREIRA, D. R. M. *et al.* Um modelo de risco proporcional dependente do tempo. Universidade Federal de São Carlos, 2007. Citation on page 36.

PENG, Y.; TAYLOR, J. M.; YU, B. A marginal regression model for multivariate failure time data with a surviving fraction. **Lifetime data analysis**, Springer, v. 13, n. 3, p. 351–369, 2007. Citation on page 21.

PRICE, D. L.; MANATUNGA, A. K. Modelling survival data with a cured fraction using frailty models. **Statistics in medicine**, Wiley Online Library, v. 20, n. 9-10, p. 1515–1527, 2001. Citations on pages 21 and 61.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2013. Available: <http://www.R-project.org/>. Citations on pages 29, 38, 57, 76, and 77.

ROCHA, R.; NADARAJAH, S.; TOMAZELLA, V.; LOUZADA, F.; EUDES, A. New defective models based on the kumaraswamy family of distributions with application to cancer data sets. **Statistical methods in medical research**, SAGE Publications, p. 1–15, 2015. Citations on pages 20 and 48.

ROCHA, R.; NADARAJAH, S.; TOMAZELLA, V.; LOUZADA, F. Two new defective distributions based on the marshall–olkin extension. **Lifetime data analysis**, Springer, p. 1–25, 2015. Citations on pages 20 and 48.

ROCHA, R.; NADARAJAH, S.; TOMAZELLA, V.; LOUZADA, F.; EUDES, A. New defective models based on the kumaraswamy family of distributions with application to cancer data sets. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 26, n. 4, p. 1737–1755, 2017. Citations on pages 20 and 31.

ROCHA, R. F.; TOMAZELLA, V. L. D.; LOUZADA, F. Inferência clássica e bayesiana para o modelo de frção de cura gompertz defeituoso. **Rev. Bras. Biom**, v. 32, n. 1, p. 104–114, 2014. Citation on page 20.

ROCHA, R. F. d. Defective models for cure rate modeling. Universidade Federal de São Carlos, 2016. Citation on page 86.

RODRIGUES, J.; CANCHO, V. G.; CASTRO, M. de; LOUZADA-NETO, F. On the unification of long-term survival models. **Statistics and Probability Letters**, Elsevier, v. 79, n. 6, p. 753–759, 2009. Citations on pages 19, 25, and 26.

SANTOS, M. R.; ACHCAR, J. A.; MARTINEZ, E. Z. Bayesian and maximum likelihood inference for the defective gompertz cure rate model with covariates. **Ciência e Natura**, Universidade Federal de Santa Maria-Centro de Ciências Naturais e Exatas, v. 39, n. 2, p. 244, 2017. Citation on page 20.

SCHWARZ, G. *et al.* 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 30.

SCUDILIO, J.; CALSAVARA, V. F.; ROCHA, R.; LOUZADA, F.; TOMAZELLA, V.; RODRIGUES, A. S. Defective models induced by gamma frailty term for survival data with cured fraction. **Journal of Applied Statistics**, Taylor & Francis, v. 46, n. 3, p. 484–507, 2019. Citations on pages 22, 45, and 86.

SINHA, D.; DEY, D. K. Semiparametric bayesian analysis of survival data. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 92, n. 439, p. 1195–1212, 1997. Citation on page 21.

TOMAZELLA, V. L. D. **Modelagem de dados de eventos recorrentes via processo de Poisson com termo de fragilidade**. Phd Thesis (PhD Thesis) — Instituto de Ciências Matemáticas e de Computação, 2003. Citations on pages 27 and 49.

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 20, 21, 27, and 49.

WHITMORE, G. A. An inverse gaussian model for labour turnover. **Journal of the Royal Statistical Society. Series A (General)**, JSTOR, p. 468–478, 1979. Citation on page 20.

WIENKE, A. Frailty models. **Wiley StatsRef: Statistics Reference Online**, Wiley Online Library, 2003. Citations on pages 28 and 50.

YAU, K. K.; NG, A. S. Long-term survivor mixture model with random effects: application to a multi-centre clinical trial of carcinoma. **Statistics in medicine**, Wiley Online Library, v. 20, n. 11, p. 1591–1607, 2001. Citation on page 21.

YU, B.; PENG, Y. Mixture cure models for multivariate survival data. **Computational Statistics & Data Analysis**, Elsevier, v. 52, n. 3, p. 1524–1532, 2008. Citation on page 21.

ZOGRAFOS, K.; BALAKRISHNAN, N. On Families of Beta- and Generalized Gamma-Generated Distributions and Associated Inference. **Statistical Methodology**, v. 6, p. 344–362, 2009. Citation on page 21.

APPENDIX A

BREAST CANCER DATA SET

Table 14 – Breast cancer data set collected at the hospital A.C.Camargo Cancer Center.

ID	Time	Censored	TIL	N	T	ID	Time	Censored	TIL	N	T
1	36.73	1	-	1	1	40	62.50	0	1	1	1
2	162.87	0	-	1	0	41	62.13	0	1	1	1
3	160.07	0	1	0	0	42	27.43	1	0	1	0
4	135.40	0	0	1	-	43	30.17	0	0	1	1
5	125.13	0	0	0	1	44	41.27	1	0	1	1
6	128.80	0	-	0	1	45	29.87	0	1	1	1
7	119.37	0	-	0	0	46	20.37	1	0	1	1
8	72.43	0	1	1	0	47	28.90	1	0	0	1
9	110.60	0	-	0	1	48	20.70	1	0	1	0
10	84.50	0	-	1	1	49	54.53	0	0	1	0
11	20.83	1	0	1	1	50	43.40	0	0	1	1
12	25.93	1	-	1	1	51	44.73	0	-	1	0
13	98.77	0	-	0	1	52	50.40	0	0	1	1
14	90.63	0	0	1	1	53	48.63	0	0	1	0
15	40.47	0	1	1	1	54	11.53	1	1	1	1
16	9.10	1	0	1	1	55	52.43	0	1	1	0
17	9.00	1	0	1	1	56	50.57	0	-	0	0
18	35.47	1	-	1	1	57	49.90	0	1	1	0
19	30.17	0	-	1	1	58	41.90	0	1	1	1
20	100.53	0	0	1	1	59	16.50	1	0	1	1
21	16.27	1	0	1	0	60	42.93	0	0	0	1
22	67.87	0	1	1	1	61	44.77	0	0	0	0
23	11.63	1	-	1	1	62	35.90	0	1	0	0
24	84.00	0	0	1	1	63	38.87	0	1	1	1
25	22.70	1	1	1	0	64	28.03	0	0	1	0
26	78.90	0	0	1	1	65	32.43	0	0	0	1
27	59.93	0	-	1	0	66	26.20	0	0	0	1
28	60.77	1	1	1	1	67	31.03	0	-	1	1
29	24.40	1	0	1	1	68	32.77	0	0	0	0
30	77.13	0	1	0	1	69	30.63	0	1	0	0
31	36.67	1	0	1	0	70	32.00	1	-	1	1
32	17.60	0	-	1	0	71	29.33	0	0	1	0
33	59.33	0	-	1	1	72	28.20	0	0	0	0
34	66.03	0	0	1	0	73	26.37	0	-	0	0
35	63.83	0	1	1	0	74	27.53	0	1	1	1
36	63.50	0	1	0	0	75	18.77	1	1	1	1
37	64.57	0	1	1	1	76	24.43	0	1	1	0
38	20.53	1	0	1	1	77	41.70	1	-	0	1
39	14.77	1	0	0	1	78	47.27	1	0	1	1