Modelo Logístico Generalizado Dependente do Tempo com Fragilidade

Eder Angelo Milani

Modelo Logístico Generalizado Dependente do Tempo com Fragilidade

Eder Angelo Milani

Orientadora: Profª. Dra. Vera L. D. Tomazella

Coorientadora: Profª. Dra. Teresa Cristina Martins Dias

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

UFSCar - São Carlos

Fevereiro/2011

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

M637ml

Milani, Eder Angelo.
Modelo logístico generalizado dependente do tempo com fragilidade / Eder Angelo Milani. -- São Carlos : UFSCar, 2011.
74 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2011.

1. Estatística. 2. Inferência bayesiana. 3. Análise de sobrevivência. 4. Fragilidade. 5. Modelo logístico generalizado dependente do tempo. 6. Probabilidade de cobertura. I. Título.

CDD: 519.5 (20ª)

Eder Angelo Milani

MODELO LOGÍSTICO GENERALIZADO DEPENDENTE DO TEMPO COM FRAGILIDADE

Dissertação apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovada em 11 de fevereiro de 2011.

BANCA EXAMINADORA

Presidente _____
Profa. Dra. Vera Lucia Damasceno Tomazella (DEs-UFSCar/Orientadora)

1° Examinador _____
Prof. Dr. Enrico Antonio Colosimo (UFMG)

2° Examinador _____
Prof. Dr. Francisco Louzada Neto (DEs-UFSCar)

3° Examinador _____
Profa. Dra. Teresa Cristina Martins Dias (DEs-UFSCar/Co-Orientadora)

Agradecimentos

Agradeço primeiramente a Deus por me dar forças para alcançar o final do trabalho.

À minha mãe, Maria, ao meu pai, José, à minha irmã, Eliana, e à minha avó, Antônia, que me deram todo o apoio e incentivo para que eu chegasse até esta etapa de minha vida.

À minha namorada, Amanda Buosi Gazon, pelas infinitas horas de conversa, pois a distância era grande e a saudade ainda maior.

À Profª. Drª. Vera L. D. Tomazella pela orientação e por confiar na minha capacidade talvez mais do que eu mesmo.

À Profª. Drª. Teresa Cristina Martins Dias por além de me orientar, se tornar uma amiga que nunca mais irei esquecer.

Aos Professores Francisco Louzada Neto e Enrico A. Colosimo pelas sábias sugestões e comentários que foram de fundamental importância para alcançar o final deste trabalho.

Aos meus amigos e colegas que me acompanharam durante a realização deste trabalho, em especial, Marcos Henrique, Vinicius, Paulo, Rafael, Rodrigo, Rubens e Marcos de Almeida.

A todos do departamento de estatística que contribuíram de maneira direta ou indireta para a minha formação.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo apoio financeiro (Processo 2009/06443-6).

Resumo

Vários autores têm preferido modelar dados de sobrevivência na presença de covariáveis por meio da função de risco, fato este relacionado à sua interpretação. A função de risco descreve como a taxa instantânea de falha se modifica com o passar do tempo. Neste contexto, um dos modelos mais utilizados é o modelo de Cox (1972) sendo que a suposição básica para o seu uso é que a razão das taxas de falhas, de dois quaisquer indivíduos, sejam proporcionais. Contudo, experiências mostram que existem dados de sobrevivência que não podem ser acomodados pelo modelos de Cox. Este fato tem sido determinante no desenvolvimento de vários tipos de modelos de risco não proporcional. Entre eles podemos citar o modelo de falha acelerado (Prentice, 1978), o modelo de risco híbrido (Etezadi-Amoli e Ciampi, 1987) e os modelos de risco híbrido estendido (Louzada-Neto, 1997 e 1999). Mackenzie (1996) propôs uma nova família paramétrica de modelo de risco não proporcional intitulado modelo de risco logístico generalizado dependente do tempo (Generalized time-dependent logistic model-GTDL). Este modelo é baseado na generalização da função logística padrão para a forma dependente do tempo e é motivado em parte por considerar o efeito do tempo em seu ajuste e, em parte pela necessidade de considerar estrutura paramétrica. O modelo de fragilidade (Vaupel *et al.*, 1979, Tomazella, 2003, Tomazella *et al.*, 2004) é caracterizado pela utilização de um efeito aleatório, ou seja, de uma variável aleatória não observável, que representa as informações que não podem ou não foram observadas, como por exemplo, fatores ambientais e genéticos, ou ainda informações que, por algum motivo, não foram consideradas no planejamento. A variável de fragilidade é introduzida na modelagem da função de risco, com o objetivo de controlar a heterogeneidade não observável das unidades em estudo, inclusive a dependência das unidades que compartilham os mesmos fatores de risco. Neste trabalho consideramos uma extensão do modelo GTDL utilizando o modelo de fragilidade como uma alternativa para

modelar dados que não tem uma estrutura de risco proporcional. Sob uma perspectiva Clássica, fizemos um estudo de simulação e uma aplicação com dados reais. Também utilizamos a abordagem Bayesiana para um conjunto de dados reais.

Palavras-chave: Abordagem bayesiana. Análise de sobrevivência. Fragilidade. Modelo logístico generalizado dependente do tempo. Probabilidade de cobertura.

Abstract

Several authors have preferred to model survival data in the presence of covariates through the hazard function, a fact related to its interpretation. The hazard function describes as the instantaneous average of failure changes over time. In this context, one of the most used models is the Cox's model (1972), in which the basic supposition for its use is that the ratio of the failure rates, of any two individuals, are proportional. However, experiments show that there are survival data which can not be accommodated by the Cox's model. This fact has been determinant in the developing of several types of non-proporcional hazard models. Among them we mention the accelerated failure model (Prentice, 1978), the hybrid hazard model (Etezadi-Amoli and Ciampi, 1987) and the extended hybrid hazard models (Louzada-Neto, 1997 and 1999). Mackenzie (1996) proposed a parametric family of non-proportional hazard model called generalized time-dependent logistic model - GTDL. This model is based on the generalization of the standard logistic function for the time-dependent form and is motivated in part by considering the time-effect in its setting and, in part by the need to consider parametric structure. The frailty model (Vaupel et al., 1979, Tomazella, 2003, Tomazella et al., 2004) is characterized by the use of a random effect, ie, an unobservable random variable, which represents information that or could not or were not collected, such as, environmental and genetics factors, or yet information that, for some reason, were not considered in the planning. The frailty variable is introduced in the modeling of the hazard function, with the objective of control the unobservable heterogeneity of the units under study, including the dependence of the units that share the same hazard factors. In this work we considered an extension of the GTDL model using the frailty model as an alternative to model data which does not have a proportional hazard structure. From a classical perspective, we did a simulation study and an application with real data. We also used a Bayesian approach to a real data set.

Keywords: Bayesian pproach. Frailty. Generalized time-dependent logistic model. Probability of coverage. Survival analysis.

Sumário

Capítulo 1

Introdução

A teoria para dados de sobrevivência tem sido bastante desenvolvida com o objetivo de estudar a função de risco/sobrevivência de um indivíduo ou sistema.

A análise de sobrevivência permite determinar quais variáveis afetam a forma da função de risco e obter estimativas destas funções para cada indivíduo ou componente. Este estudo envolve o acompanhamento de indivíduos até a ocorrência de algum evento de interesse, por exemplo, a falha (morte) do mesmo. Em estudos médicos, uma abordagem padrão para analisar dados de sobrevivência é empregar o modelo de riscos proporcionais de Cox (1972). Essa abordagem permite incorporar covariáveis e estimar os parâmetros de regressão associados a estas covariáveis.

Cada vez mais experiências têm mostrado que nem todos dados de sobrevivência comportam a suposição de risco proporcional, o que tem conduzido para o desenvolvimento de vários tipos de modelos de riscos não proporcionais. Como exemplo citamos: modelo de riscos aditivos para dados agrupados (Aranda-Ordaz, 1983 e Tibshirani e Ciampi, 1983), modelo de risco logístico generalizado dependente do tempo (Mackenzie, 1996), que foi utilizado por Mackenzie (1997) para eventos recorrentes e modelo parcialmente paramétrico de McKeague e Sasieni (1994). Este último modelo explora o desenvolvimento moderno da teoria de processos de contagem aplicados a estrutura semiparamétrica discutida por Andersen e Gill (1982). Contudo o uso de modelos paramétricos que não apresentam a estrutura de risco proporcionais oferecem uma abordagem alternativa, a qual não podem ser descartada.

Uma suposição usual feita em análise de efeito de tratamento ou fatores de risco em sobrevivência é que indivíduos são condicionalmente independentes dada as covariáveis observadas. Quando se trata de eventos múltiplos ou repetitivos para um mesmo indivíduo a suposição de dependência pode ser questionável. Motivado por esta situação de dependência Vaupel *et al.* (1979) introduziu o primeiro modelo de fragilidade.

A variável de fragilidade é introduzida na modelagem da função de risco com o objetivo de controlar a heterogeneidade não observável das unidades em estudo inclusive a dependência, representando as informações que ou não podem ou não foram observadas, tais como fatores ambientais e genéticos, ou informações que por algum motivo não foram consideradas no planejamento.

Suponha que dois indivíduos tenham os mesmos valores das covariáveis observadas, mas nem por isso é esperado que estes venham a experimentar o evento de interesse ao mesmo tempo. Neste caso existem fatores genéticos e ambientais, entre outros, que ou não podem ser medidos ou que por algum motivo não foram incorporados no estudo, mas que influenciam na determinação do tempo de falha. Esta é uma situação, dentre outras, que é mais interessante utilizar modelos com fragilidade.

Nos modelos de fragilidade as candidatas naturais à distribuição de fragilidade, devido as suas características, são as distribuições gama, log-normal, inversa gaussiana e Weibull. Vaupel *et al.* (1979) e Clayton (1978) foram os primeiros a usar modelos de fragilidade considerando a distribuição gama; Oakes (1982) foi o primeiro a considerar o modelo de fragilidade para dados multivariados, usando também a gama para a variável de fragilidade. Clayton e Cuzick (1985) e Tomazella *et al.* (2004) são alguns dos autores que utilizaram a distribuição gama em seus estudos. Outras distribuições foram propostas, como por exemplo, Hougaard (1984) sugeriu a distribuição gaussiana inversa para estudar métodos de tabela de vida, Korsgaard *et al.* (1998) utilizou a distribuição log-normal, Hougaard (1986a, b) sugeriu uma distribuição que generaliza, como por exemplo, a gama e a gaussiana inversa e Aalen (1988), propôs uma distribuição que tem como caso particular a distribuição sugerida por Hougaard (1986a, b).

Neuhaus *et al.* (1992) e Henderson e Oman (1999) mostraram que ignorar a fragilidade, quando esta existe, pode levar a vícios na estimação do efeito das covariáveis, e com isso levar a erros na interpretação dos coeficientes do modelo de regressão.

Com o avanço e popularização de computadores e *softwares* os cálculos mais rápidos e precisos beneficiaram a utilização da abordagem bayesiana, Clayton (1991) foi o primeiro autor a utilizar a abordagem bayesiana para modelos com fragilidade, outros depois também usaram esta abordagem, dentre eles podemos citar Korsgaard *et al.* (1998), Tomazella *et al.* (2004).

Na Seção 1.1 apresentamos uma breve introdução sobre análise de sobrevivência, com alguns conceitos introdutórios como a construção da verossimilhança para dados censurados e uma técnica não-paramétrica para estimar a função de sobrevivência. Na Seção 1.2 apresentamos a modelagem via função de risco utilizando o modelo de Cox. O objetivo do nosso trabalho está apresentado na Seção 1.3.

1.1 Introdução à Análise de Sobrevivência

Quando estudamos o tempo de vida de determinado componente (indivíduo), e queremos saber quando este vem a falha (morrer), notamos que o tempo em que ocorre o evento não é determinístico, ou seja, trabalhamos com um evento aleatório representado pelo tempo de ocorrência. Para estudarmos o evento, suponhamos T uma variável aleatória que represente o tempo até a falha do componente (morte do indivíduo). A área da estatística que estuda esta situação é chamada análise de confiabilidade (sobrevivência).

Em análise de sobrevivência os dados são caracterizados pelos tempos de falha e, muito freqüentemente, pelas censuras. Normalmente um conjunto de covariáveis também é adotado no estudo.

Uma dúvida que geralmente se apresenta para estudantes que iniciam na área de análise de sobrevivência é: por que não utilizar métodos de regressão usuais para analisar estes dados?

A resposta é muito simples: modelos de regressão usuais não incluem o fato de termos dados censurados. Como, em geral, o tamanho da amostra em estudos clínicos não é grande, descartar os dados censurados neste caso seria inviável. Com isso, precisamos desenvolver métodos para analisar observações que apresentam censuras.

Suponha que estamos interessados em estudar uma determinada doença, ou melhor dizendo, estudarmos o tempo que alguns pacientes demoram para morrer devido a uma determinada doença. Suponha também que o estudo (coleta de informações) tem um limite de tempo e, existem indivíduos que estão no estudo e não chegaram a morrer neste intervalo de tempo e ainda não queremos descartar estes indivíduos do estudo. Conseqüentemente adotaremos que o tempo de falha dos indivíduos que não morreram no intervalo de tempo do estudo é maior do que o tempo máximo de estudo. Dizemos então, que estes tempos são censurados. Adicionaremos então ao estudo uma variável que indicará se os tempos observados são de falha ou de censura.

A seguir definiremos as censuras mais comuns em estudos de análise de sobrevivência.

• Censura Tipo I ou à direita

Este tipo de censura ocorre quando, geralmente, o tempo para o fim do estudo é pré-estabelecido; assim, alguns indivíduos deixam de experimentar o evento de interesse ao fim deste estudo, tendo os seus tempos de vida censurados à direita. Um exemplo para esse tipo de censura é quando um determinado banco deseja verificar o tempo até que os clientes, de determinada carteira, se tornam inadimplentes. Estuda-se portanto, esta carteira durante um tempo pré-determinado pela instituição e ao fim, alguns desses deixaram de experimentar o evento de interesse (portanto não são inadimplentes), observando assim, a censura do tipo I.

• Censura Tipo II

Quando o estudo será terminado após um determinado número, n, de indivíduos experimentar o evento de interesse, ou seja, após um número n de ocorrências a pesquisa é finalizada e os indivíduos que deixaram de experimentar o evento de interesse terão seus tempos censurados.

• Censura Aleatória

Diferentemente das outras, este tipo de censura foge ao controle do pesquisador. Geralmente ocorre quando o indivíduo abandona determinado experimento sem ter experimentado o evento de interesse. A censura aleatória é um caso mais comum, tendo como caso particular a censura tipo I ou censura à direita, por exemplo, se o paciente

morrer por uma razão diferente da estudada.

Existem outros tipos de censuras (ver por exemplo Lawless, 1982, Colosimo e Giolo, 2006).

Um fato importante é identificar se o tempo observado é tempo de falha ou de censura; para isso é adotado uma variável que é definida da seguinte forma,

$$\delta_i = \begin{cases} 1, & \text{se é tempo de falha} \\ 0, & \text{se é um tempo de censura,} \end{cases} \tag{1.1}$$

sendo que $i = 1, \ldots, n$ representa as n unidades (indivíduos).

Suponhamos que a variável aleatória T, $T \geq 0$, tenha função de densidade de probabilidade denotada por $f(t)$. Podemos descrever a função de densidade de probabilidade como o limite da probabilidade de um indivíduo falhar no intervalo de tempo $[t, t + \Delta t]$ por unidade de tempo e expressar como sendo,

$$f(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}. \tag{1.2}$$

A estimação da probabilidade de um indivíduo sobreviver pelo menos até o tempo t, é um dos principais interesse na análise de sobrevivência. Para estimar esta probabilidade definimos a função de sobrevivência, que é dada por,

$$S(t) = P(T \geqslant t) = 1 - P(T < t) = 1 - \int_0^t f(u)du = 1 - F(t), \tag{1.3}$$

sendo $F(t)$ a função distribuição acumulada.

A função (1.3) é decrescente no intervalo de tempo $[0, \infty)$, tal que $S(0) = 1$ e $S(\infty) = 0$.

O taxa instantâneo de um indivíduo sofrer o evento no intervalo $[t, t + \Delta t]$, dado que ele sobreviveu até o tempo t é dado por

$$h(t) = \lim_{\Delta t \longrightarrow 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{1.4}$$

Existem algumas importantes relações entre as equações (1.2), (1.3) e (1.4) que são

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

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\left(lnS(t)\right), \tag{1.5}$$

$$S(t) = e^{-\int_0^t h(u)du} = e^{-H(t)}. \tag{1.6}$$

.

Algumas outras relações ainda podem ser construídas a partir destas (ver Colosimo e Giolo, 2006).

1.1.1 Função de Verossimilhança

Como é comum em análise de sobrevivência a presença de observações censuradas, consideramos tal fato na hora de construir a função de verossimilhança.

Seja uma amostra aleatória de tamanho n na qual observamos, para cada indivíduo i $(i = 1, \ldots, n)$, a seguinte dupla (t_i, δ_i), temos como variável resposta o tempo (t_i) e δ_i informando se o tempo observado é de falha ou não.

Sabemos que no caso de observações não censuradas, a contribuição para a verossimilhança é a função de densidade de probabilidade. Para observações censuradas consideramos que a contribuição desta será a função de sobrevivência. Vamos supor que temos um modelo paramétrico com apenas um parâmetro, denotado por θ e considerando a censura não informativa. Com isso obtemos a função de verossimilhança para θ, dada por,

$$L(\theta|\mathbf{t}, \boldsymbol{\delta}) = \prod_{i=1}^{n}[f(t_i;\theta)]^{\delta_i}[S(t_i;\theta)]^{1-\delta_i}, \tag{1.7}$$

utilizando a equação (1.5) reescrevemos a equação (1.7) da seguinte forma

$$L(\theta|\mathbf{t}, \boldsymbol{\delta}) = \prod_{i=1}^{n}[h(t_i;\theta)]^{\delta_i}[S(t_i;\theta)]. \tag{1.8}$$

A função de verossimilhança dada pela equação (1.8) é válida para todos os tipos de censuras citadas neste trabalho.

1.1.2 Técnicas Não-Paramétricas

Na literatura de análise de sobrevivência são encontrados alguns estimadores da função de sobrevivência obtidos por técnicas não-paramétricas. Podemos citar, como exemplo, o estimador de Nelson-Aalen proposto por Nelson (1972) e o proposto por Kaplan e Meier (1985). Este último é um dos mais utilizados e por isso apresentamos a seguir.

Estimador de Kaplan-Meier

O estimador de Kaplan-Meier sugerido por Kaplan e Meier (1985) é uma adaptação da função de sobrevivência empírica, pois este incorpora dados censurados. A função de sobrevivência empírica, na ausência de censuras, é definida como

$$\widehat{S_E}(t) = \frac{\text{nº de observações que não falharam até o tempo t}}{\text{nº total de observações no estudo}},$$

sendo $\widehat{S_E}(t)$ uma função escada com degraus nos tempos observados de falha de tamanho $1/n$, e n o tamanho da amostra. Se existir empates em certo tempo t, o tamanho do degrau fica multiplicado pelo número de empates.

O estimador de Kaplan-Meier, na sua construção, considera tantos intervalos de tempo quantos forem o número de falhas distintas. Ou seja, suponha n indivíduos no estudo e k falhas distintas nos tempos $t_1 < t_2 < \ldots < t_k$. Considerando $S_{KM}(t)$ uma função de probabilidade discreta com probabilidade diferente de zero somente nos tempos de falha t_j, $j = 1, \ldots, k$, tem-se que

$$S_{KM}(t_j) = (1 - q_1)(1 - q_2) \ldots (1 - q_j),$$

sendo que q_j é a probabilidade de um indivíduo morrer no intervalo $[t_{j-1}, t_j)$ sabendo que ele não morreu até t_{j-1}. Considere $t_0 = 0$.

O estimador de Kaplan-Meier se reduz a estimar q_j que é dado por:

$$\widehat{q_j} = \frac{\text{nº de falhas em } t_j}{\text{nº de observações sob risco em } t_{j-1}}$$

para $j = 1, ..., k$.

A expressão geral do estimador de Kaplan-Meier pode ser apresentada da seguinte forma,

$$\widehat{S_{KM}}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right)$$

sabendo que,

- $t_1 < t_2 < ... < t_k$, para k tempos distintos e ordenados de falhas,
- d_j é o número de falhas em t_j, $j = 1, ..., k$ e
- n_j é o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j, $j = 1, ..., k$ (Colosimo e Giolo, 2006).

1.2 Modelagem via Função de Risco

Quando estamos interessados em estudar a sobrevivência de pacientes, uma alternativa é usar a função de risco. O primeiro modelo proposto na literatura que utiliza covariáveis e a modelagem de risco é o de Cox (1972) e desde então vem sendo extremamente utilizado. Este fato pode ser explicado pela seguinte frase "Este modelo é o mais utilizado em estudos clínicos por sua versatilidade "(ver Colosimo e Giolo, 2006), a versatilidade deste modelo é devido aos fatos que a função de risco pode assumir diversas formas e a forma semi-paramétrico do modelo deixa-lo flexível.

1.2.1 Modelo de Riscos Proporcionais de Cox

Considere que para n indivíduos foram coletados o seguinte termo $(t_i, \delta_i, \mathbf{x}_i)$, sendo t_i e δ_i como apresentados em (1.1) e $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $i = 1, ..., n$ e p é a quantidade de covariáveis. O modelo de risco proporcional proposto por Cox (1972) é dado por

$$h(t) = h_0(t) g(\mathbf{x}'\boldsymbol{\beta}),$$

sendo que $h_0(t)$ representa a função de risco de base para a unidade quando $\mathbf{x} = 0$, $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ é o vetor de coeficientes a serem estimados e $g(\mathbf{x}'\boldsymbol{\beta})$ é uma função não-negativa normalmente utilizada como $\exp(\mathbf{x}'\boldsymbol{\beta})$.

O modelo em questão pode ser chamado de paramétrico ou semi-paramétrico, pois é composto pelo produto de dois componentes, um paramétrico ($\exp(\mathbf{x}'\boldsymbol{\beta})$) e o outro ($h_0(t)$) podendo ser paramétrico ou não. O componente $h_0(t)$ quando não paramétrico é especificado ser uma função não negativa do tempo; quando paramétrico, pode assumir distribuições como Weibull, exponencial entre outras (neste caso não é mas o modelo proposto por Cox (1972), mas sim uma modificação). A outra parte, $\exp(\mathbf{x}'\boldsymbol{\beta})$, mede o efeito das covariáveis.

Chamamos o modelo de Cox de modelo de risco proporcional, pois, considerando dois indivíduos, i e j, a razão das funções de risco é

$$\frac{\lambda(t)\exp(\beta x_i)}{\lambda(t)\exp(\beta x_j)} = \frac{\exp(\beta x_i)}{\exp(\beta x_j)} = \exp[\beta(x_i - x_j)].$$

observe que o efeito do tempo desaparece e a razão é proporcional para qualquer instante de tempo t.

Este modelo assume que o vetor de covariáveis tem um efeito multiplicativo na função de risco. Este fato implica que a estrutura deste modelo impõe proporcionalidade entre as funções de risco de diferentes níveis de covariáveis, não permitindo que elas se cruzam e dependa do tempo.

Para interpretar a razão das funções de risco serem proporcional, suponha que o indivíduo i tenha o valor da sua função de risco k vezes o valor da função de risco do indivíduo j no começo do estudo, ou seja, a razão das funções de risco é igual a k. Esta razão com o passar do tempo não ira se modificar, com isso, a razão dessas mesmas funções de risco, mas agora em qualquer instante de tempo t, continua igual a k.

1.3 Objetivo

A modelagem via função de risco utilizando covariáveis é uma técnica na análise de sobrevivência que é utilizada, devido ao fato da sua interpretação. Neste contexto já citamos o modelo de Cox (1972), mas este se limita ao fato de que a razão das funções de risco de dois indivíduos é proporcional.

Vários outros modelos existentes na literatura vão de encontro ao fato de proporcionalidade, ou seja, a razão das funções de risco de dois indivíduos não são proporcionais.

Podemos citar como exemplo o modelo proposto por Mackenzie (1996) que é conhecido como modelo de risco logístico generalizado dependente do tempo (GTDL).

Usualmente fazemos a suposição que tempos de sobrevivência de indivíduos distintos são condicionalmente independentes dadas as covariáveis observadas. Em alguns casos esta suposição pode não ser válida, como por exemplo, o tempo de sobrevivência quando observado em gêmeos ou em indivíduos da mesma família. Nestas situações é esperado que o tempo de sobrevivência tenha uma certa semelhança.

Uma alternativa para modelar dados de sobrevivência quando a suposição de tempos independentes pode não ser válida, são os modelos de fragilidade. Vaupel *et al.* (1979) e Tomazella *et al.* (2004) entre outros, utilizaram modelos com fragilidade para contornar a falta de independência.

A partir do modelo GTDL e das técnicas para incorporar a fragilidade, apresentamos o modelo GTDL com fragilidade, com o intuito de desenvolver uma alternativa para modelar dados de sobrevivência quando existe heterogeneidade entre os indivíduos e a suposição de risco proporcional não se sustenta.

Utilizando inferência Clássica apresentamos em três situações (sem censura, com 10% e 30% de censura) o cálculo da probabilidade de cobertura, o vício e o erro quadrático médio. Também, duas aplicações com o modelo GTDL com fragilidade, uma com dados reais e outra com dados simulados.

Desenvolvemos uma abordagem Bayesiana para o modelo GTDL com fragilidade, aplicamos esta metodologia em dados reais e em dados simulados, sendo que os métodos MCMC serão utilizados para estudar os parâmetros de interesse.

Este trabalho está organizado da seguinte maneira: no Capítulo 2 apresentamos o modelo GTDL, com as suas propriedades, casos particulares e uma aplicação. No Capítulo 3 mostramos o modelo de fragilidade multiplicativo, o modelo GTDL com fragilidade e uma aplicação. São apresentados no Capítulo 4, um método de geração de dados do modelo GTDL com fragilidade, a probabilidade de cobertura, o vício, o erro quadrático médio, os resultados dos ajustes utilizando os modelos GTDL e GTDL com fragilidade e uma comparação entre eles. Apresentamos no Capítulo 5 uma abordagem Bayesiana para o modelo GTDL com fragilidade com uma aplicação. Para um conjunto de dados reais

ajustamos os modelos GTDL e GTDL com fragilidade e selecionamos o melhor modelo; com o modelo GTDL com fragilidade fizemos duas abordagens. No Capítulo 6 são apresentadas algumas conclusões e perspectivas de continuidade do trabalho. No Apêndice A mostramos detalhes do ajuste utilizando dados reais. O algoritmo de Metropolis-Hastings é apresentado no Apêndice B. No Apêndice C é mostrado o método de convergência de Geweke.

Capítulo 2

Modelo de Risco Logístico Generalizado Dependente do Tempo

A motivação para o desenvolvimento de modelos de risco não proporcional é devido ao interesse de explicar os conjuntos de dados que não sustentam a suposição de risco proporcional, neste contexto Mackenzie (1996) propôs o modelo de risco logístico generalizado dependente do tempo (GTDL), sendo que Cremasco (2005) apresentou um estudo de simulação e uma abordagem bayesiana para este mesmo modelo. A formulação do modelo GTDL é apresentado na Seção 2.1. Na Seção 2.2 mostramos as propriedades e casos particulares e uma aplicação com um conjunto de dados reais é apresentado na Seção 2.3.

2.1 Modelo Logístico Generalizado Dependente do Tempo

Denotando por T uma variável aleatória não negativa representando o tempo de falha, a função de risco do modelo GTDL é dado por

$$h(t|\lambda,\alpha,\boldsymbol{\beta}) = \lambda\frac{\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}, \tag{2.1}$$

sendo que $\lambda > 0$ é um escalar, α é uma medida do efeito do tempo e $\boldsymbol{\beta}'=(\beta_1,\ldots,\beta_p)$ é um vetor de p parâmetros desconhecidos medindo a influência das p covariáveis $\mathbf{x}' = (\mathbf{x}_1,\ldots,\mathbf{x}_p)$.

Também considerando T uma variável aleatória não negativa representando o

tempo de falha, a função de risco do modelo TDL (logístico dependente do tempo) é dado por

$$h_{TDL}(t|\alpha,\boldsymbol{\beta}) = \frac{\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}{1+\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}, \tag{2.2}$$

cujo α é uma medida do efeito do tempo e $\boldsymbol{\beta}'=(\beta_0,\beta_1,\ldots,\beta_p)$ é um vetor de $p+1$ parâmetros desconhecidos medindo a influência das p covariáveis $\mathbf{x}'=(1,\mathbf{x}_1,\ldots,\mathbf{x}_p)$ e β_0 é o intercepto.

Mackenzie (2002) mostra o porquê da retirada do parâmetro β_0 do modelo GTDL e em que sentido foi a evolução do modelo TDL para o modelo GTDL. A reprodução das explicações é a seguinte. De (2.1) temos,

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

e

$$\frac{\partial h(t|\lambda,\alpha,\boldsymbol{\beta})}{\partial\beta_0} = \lambda[\vartheta(t|\mathbf{x},\alpha,\boldsymbol{\beta}) - \vartheta^2(t|\mathbf{x},\alpha,\boldsymbol{\beta})].$$

Como $0<\vartheta(t|\mathbf{x},\alpha,\boldsymbol{\beta})<1$ e normalmente é pequeno, $\vartheta^2(t|\mathbf{x},\alpha,\boldsymbol{\beta})\approx 0$ então temos a seguinte relação

$$\lambda\frac{\partial h(t|\lambda,\alpha,\boldsymbol{\beta})}{\partial\lambda} \approx \frac{\partial h(t|\lambda,\alpha,\boldsymbol{\beta})}{\partial\beta_0}, \forall t>0$$

mostrando que não existe nova informação contida em β_0 e como os papéis dos parâmetros λ e β_0 são *inter-changeable*, então apenas um parâmetro é necessário no modelo.

A idéia para a original extensão do modelo TDL para o modelo GTDL foi baseada na necessidade de remover a inconveniente restrição sobre a função do risco imposta pelo modelo TDL, que é $0<h_{TDL}(t|\alpha,\boldsymbol{\beta},\mathbf{x})\leqslant 1,\forall t$. Isto levou à incorporação de $\lambda>0$ na função do risco e conseqüentemente à retirada do β_0 do modelo.

Utilizando a relação dada em (1.5), a função de sobrevivência é dada por,

$$S(t|\lambda,\alpha,\boldsymbol{\beta}) = \left\{\frac{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right\}^{-\lambda/\alpha},$$

usando a equação (1.5) temos que a função de densidade de probabilidade é dada por

$$f(t|\lambda,\alpha,\boldsymbol{\beta}) = \left(\lambda\frac{\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}\right)\left(\frac{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)^{-\lambda/\alpha}. \tag{2.3}$$

A função de verossimilhança considerando dados censurados é dada por

$$L(\alpha, \boldsymbol{\beta}, \lambda|\text{dados}) = \prod_{i=1}^{n} \left(\lambda \frac{\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})} \right)^{\delta_i} \left(\frac{1 + \exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})} \right)^{-\lambda/\alpha} \tag{2.4}$$

Este modelo é denominado modelo de riscos não proporcionais, pois a razão entre as funções de risco de dois indivíduos com covariáveis diferentes é dada por,

$$\begin{aligned} \frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_j)} &= \frac{\lambda \exp(\alpha t + \mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\alpha t + \mathbf{x}_i'\boldsymbol{\beta})} \frac{1 + \exp(\alpha t + \mathbf{x}_j'\boldsymbol{\beta})}{\lambda \exp(\alpha t + \mathbf{x}_j'\boldsymbol{\beta})} \\ &= \frac{1 + \exp(\alpha t + \mathbf{x}_j'\boldsymbol{\beta})}{1 + \exp(\alpha t + \mathbf{x}_i'\boldsymbol{\beta})} \exp[(\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\beta}]. \end{aligned}$$

Vemos que o efeito do tempo não desaparece e com isso a não-proporcionalidade se torna evidente.

2.2 Propriedades e Casos Particulares

Nesta Seção, apresentamos as principais propriedades do modelo GTDL e gráficos para ilustrar as funções de risco e sobrevivência. Os casos particulares também são expostos nesta Seção.

A função de sobrevivência se comporta de acordo com o valor de α. Temos que,

• para $\alpha > 0$, $S(0|\lambda, \alpha, \boldsymbol{\beta}) = 1$ e $S(\infty|\lambda, \alpha, \boldsymbol{\beta}) = 0$, ou seja, a função de sobrevivência é própria;

• para $\alpha < 0$, $S(0|\lambda, \alpha, \boldsymbol{\beta}) = 1$ e $S(\infty|\lambda, \alpha, \boldsymbol{\beta}) \neq 0$, com isso temos que a função de sobrevivência é imprópria, ou seja, quando $\alpha < 0$ temos um modelo de taxa de cura ou longa duração.

Na Figura (2.1) temos algumas formas da função de sobrevivência. Note que as relações citadas acima se verificam nesta ilustração. Neste caso consideramos o valor da covariável igual a 1 ($\mathbf{x} = 1$).

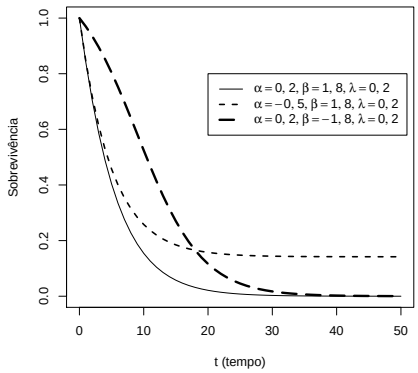

Figura 2.1: Formas da função de sobrevivência.

O comportamento da função de risco é de extremo interesse para pesquisadores que utilizam a função de risco para modelar dados de sobrevivência, o comportamento da função de risco varia de acordo com os valores de α:

- para $\alpha > 0$, a função de risco é crescente;
- para $\alpha < 0$, a função de risco é decrescente;
- para $\alpha = 0$, a função de risco é constante;

Algumas formas da função de risco e as restrições para que as funções sejam crescente e decrescente são mostradas na Figura (2.2). Aqui também utilizamos $\mathbf{x} = 1$.

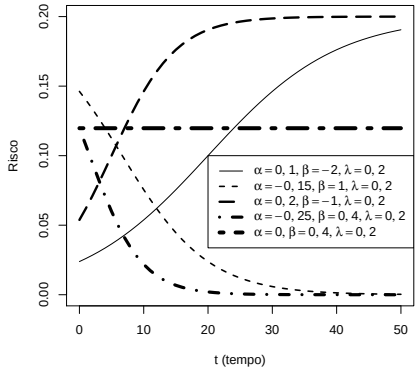

Figura 2.2: Formas da função de risco.

Notemos que o modelo dado pela equação (2.1) caracteriza uma família tri-paramétrica $(\lambda, \alpha, \boldsymbol{\beta})$ de variáveis aleatórias não negativa com função de densidade de probabilidade dada por,

$$f(t|\mathbf{x}, \lambda, \alpha, \boldsymbol{\beta}) = \lambda p(\alpha, \boldsymbol{\beta})[q(\alpha, \boldsymbol{\beta})g(\boldsymbol{\beta})]^{\lambda/\alpha},$$

sendo que os componentes individuais são funções simples do modelo logístico múltiplo dependente do tempo,

$$p(\alpha, \boldsymbol{\beta}) = \exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})/[1 + \exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})], \tag{2.5}$$

$$q(\alpha, \boldsymbol{\beta}) = 1/[1 + \exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})], \tag{2.6}$$

$$g(\boldsymbol{\beta}) = 1 + \exp(\mathbf{x}'\boldsymbol{\beta}). \tag{2.7}$$

Os casos particulares do modelo GTDL são descritos em 2.2.1 a 2.2.4.

2.2.1 Modelo de Taxa de Cura ou Longa Duração

Quando $\alpha < 0$ temos que o modelo GTDL se torna um modelo de longa duração, para mais detalhes de modelos de longa duração ver Ibrahim *et al.* (2001). Um fato importante quando se tem um modelo de longa duração é a expressão da fração de cura. Para encontrarmos esta fração fazemos o limite quando $t \longrightarrow \infty$ na função de sobrevivência; com isso, encontramos a fração de curados ou taxa de cura, que chamamos de p,

$$\begin{aligned} p = \lim_{t\longrightarrow\infty} S(t|\mathbf{x}, \lambda, \alpha, \boldsymbol{\beta}) &= \lim_{t\longrightarrow\infty} \left(\frac{1 + \exp(t\alpha + \mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}\right)^{\frac{-\lambda}{\alpha}} \\ &= \left(\frac{1}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}\right)^{\frac{-\lambda}{\alpha}} \lim_{t\longrightarrow\infty} (1 + \exp(t\alpha + \mathbf{x}'\boldsymbol{\beta}))^{\frac{-\lambda}{\alpha}} \\ &= (1 + \exp(\mathbf{x}'\boldsymbol{\beta}))^{\frac{\lambda}{\alpha}}. \end{aligned} \tag{2.8}$$

Para exemplificar este caso particular geramos um gráfico com os seguintes valores para os parâmetros

$$\alpha = -0,30;\ \lambda = 0,15 \text{ e } \beta = 3,00.$$

Adotamos que a covariável receba dois valores, desta forma simulamos dois grupos, a covariável é definida por

$$\begin{cases} x = 1, & \text{grupo A} \\ x = 0, & \text{grupo B.} \end{cases} \tag{2.9}$$

Utilizando a equação (2.8) encontramos os seguintes valores para a fração de cura

- para o grupo A, $p = 0,22$;
- para o grupo B, $p = 0,71$.

A Figura 2.3 representa as funções de sobrevivência em (a) e a de risco em (b). Na Figura 2.3 (a) existem linhas representando o valor da fração de cura para os grupos e a longa duração fica evidente. As curvas de risco apresentadas na Figura 2.3 (b) são decrescente.

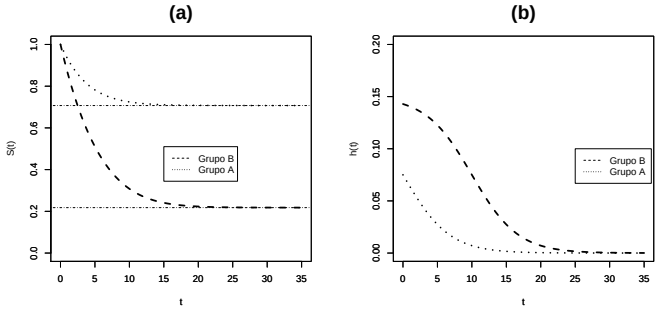

Figura 2.3: Funções de sobrevivência em (a) e de risco em (b).

A Figura 2.4 mostra algumas formas da função densidade de probabilidade dada em (2.3) com a restrição $\alpha < 0$, observamos que a curvas são decrescente.

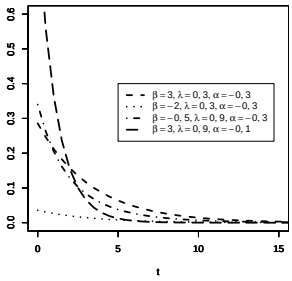

Figura 2.4: Formas da função de densidade de probabilidade.

2.2.2 Modelo Exponencial

Segundo Kalbfleisch e Prentice (1980) o modelo exponencial é dado por

$$h(t|\lambda, \boldsymbol{\beta}) = \lambda \exp(\mathbf{x}'\boldsymbol{\beta}). \tag{2.10}$$

Utilizando o modelo dado pela equação (2.1) e fazendo $\alpha = 0$, temos a seguinte função de risco

$$h(t|\lambda, \boldsymbol{\beta}) = \lambda \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}. \tag{2.11}$$

Vemos claramente que as equações (2.10) e (2.11) são diferentes, mas se observamos o comportamento das funções vemos algo em comum. O fato em comum é que as duas funções de risco são constantes, para todo tempo t, este comportamento é típico do modelo exponencial.

Temos que a função de sobrevivência, utilizando a parametrização dada pela função de risco da equação (2.11) é

$$S(t|\lambda, \boldsymbol{\beta}) = \exp\left(\frac{-\lambda t \exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}\right).$$

A função densidade de probabilidade é dada por

$$f(t|\lambda, \boldsymbol{\beta}) = \lambda \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})} \exp\left(\frac{-\lambda t \exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}\right). \tag{2.12}$$

Ilustramos o modelo exponencial adotando os seguintes valores para os parâmetros

$$\beta = 3,00;\ \lambda = 0,30 \text{ e } \alpha = 0.$$

Para simular dois grupos, a covariável x irá assumir dois valores, da mesma forma como declarado em (2.9).

Dessa forma construímos as curvas das funções de sobrevivência e do risco, estas são mostradas na Figura 2.5 em (a) e em (b), respectivamente. As curvas de risco são constante, como esperado.

Algumas formas da função de densidade de probabilidade dada em (2.12) é apresentada na Figura 2.6, observamos que as curvas são decrescente.

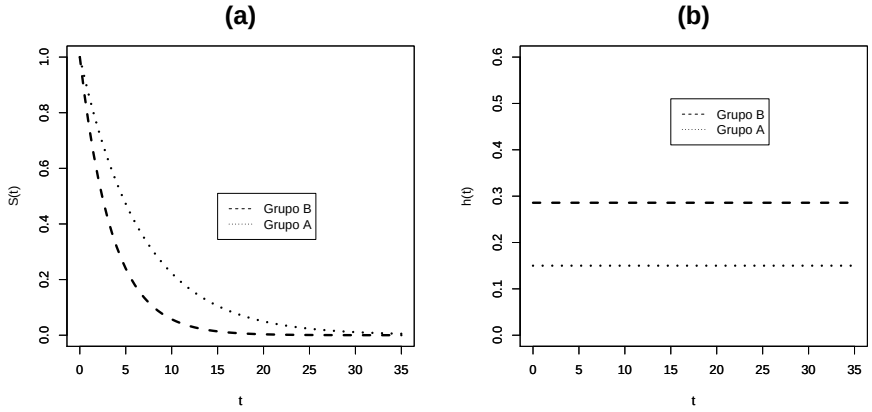

Figura 2.5: Funções de sobrevivência em (a) e de risco em (b).

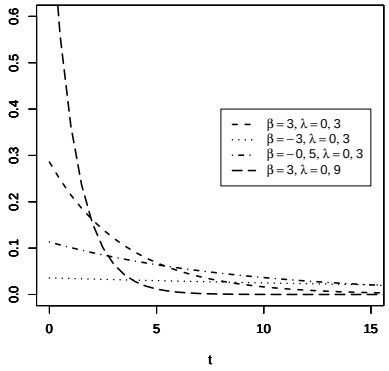

Figura 2.6: Formas da função de densidade de probabilidade.

2.2.3 Modelo de Gompertz

Gompertz (1825) sugeriu uma lei de progressão geométrica para a mortalidade de um indivíduo depois de uma certa idade, ou ainda, uma lei de mortalidade capaz de explicar como o tempo influencia na morte de um indivíduo.

A função de risco de Gompertz é dada por,

$$h(t|\phi,\eta) = \phi \exp(\eta t) \tag{2.13}$$

e ainda, a função de sobrevivência é

$$S(t|\phi,\eta) = \exp\left\{-\left(\frac{\phi}{\eta}\right)(e^{\eta t}-1)\right\}.$$

Quando

$$q(\alpha,\boldsymbol{\beta}) \to 1, \tag{2.14}$$

temos que o modelo GTDL se aproxima do modelo de Gompertz.

Utilizando a equação do risco dada em (2.1) e a restrição dada na equação (2.14) podemos escrever

$$h(t|\lambda, \alpha, \boldsymbol{\beta}) \approx \lambda \exp(\alpha t + \mathbf{x}'\boldsymbol{\beta}) = \lambda \exp(\alpha t + \gamma) = \lambda \exp(\gamma) \exp(\alpha t) \tag{2.15}$$

sendo que $\gamma = \mathbf{x}'\boldsymbol{\beta}$.

Comparando as equações (2.15) e (2.13) vemos que a função de risco dada pela equação (2.15) é uma função de risco dada pelo modelo de Gompertz com $\eta = \alpha$ e $\phi = \lambda \exp(\gamma)$.

Para exemplificar a aproximação do modelo de Gompertz para o modelo GTDL consideramos os seguintes valores para os parâmetros

$$\alpha = -0,05;\ \beta = -1,35;\ \lambda = 1,00 \text{ e } x = 1,00.$$

Utilizando os valores dos parâmetros calculamos o valor da restrição (2.14) para alguns valores de t, o resultado se encontra na Tabela 2.1.

Tabela 2.1: Valores da restrição.

Tempo	Valor
0,50	0,7982
2,00	0,8100
5,00	0,8320
10,00	0,8641
30,00	0,9453
50,00	0,9792
80,00	0,9953

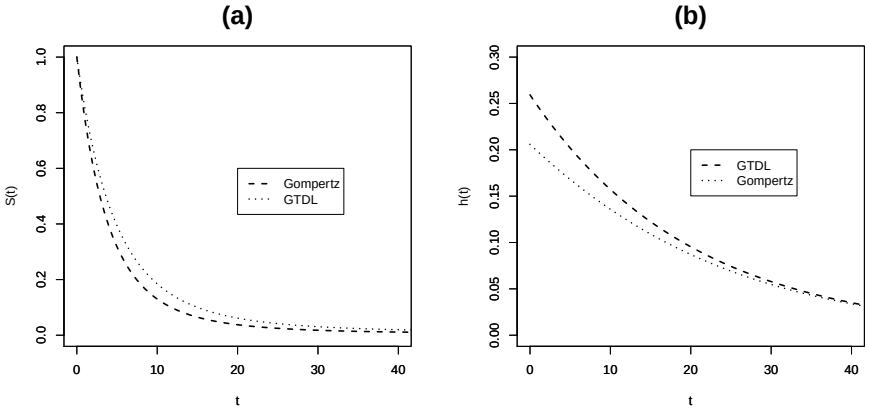

Figura 2.7: Comparação entre os modelos de Gompertz e GTDL utilizando as funções de sobrevivência em (a) e de risco em (b).

Notamos pela Tabela 2.1 que com o aumento do tempo o valor da restrição fica mais próxima de 1. A Figura 2.7 (a) é uma comparação entre as funções de sobrevivência utilizando os modelos de Gompertz e o GTDL. A comparação entre as curvas das funções de risco são apresentadas na Figura 2.7 (b). Observamos que a aproximação pode ser considerada razoável, mas quanto melhor for a aproximação da equação (2.14), melhor será a aproximação do modelo de Gompertz ao modelo GTDL.

A Figura 2.8 representa algumas funções de risco e de sobrevivência para o modelo de Gompertz. Observamos pela Figura 2.8 (a) que as curvas da função de sobrevivência podem ser com o decréscimo mais ou menos acentuado. As curvas de risco podem ser crescente ou decrescente, isto é observado pela Figura 2.8 (b).

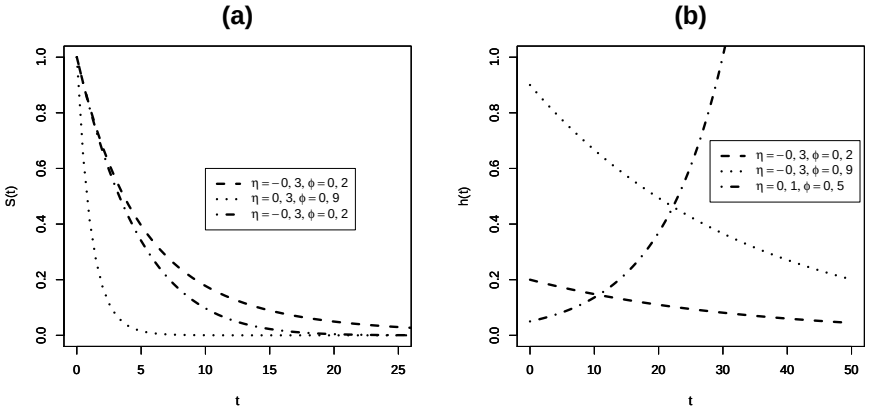

Figura 2.8: Funções de sobrevivência em (a) e de risco em (b).

Notamos pela Figura 2.9 algumas formas da função de densidade de probabilidade, as curvas são decrescente ou unimodal assimétrica.

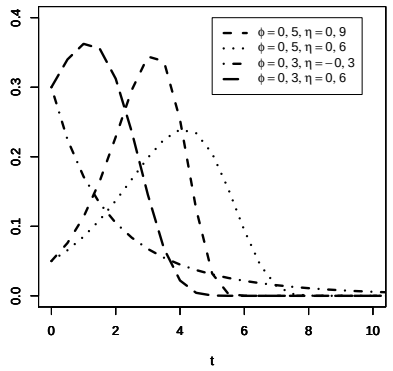

Figura 2.9: Formas da função de densidade de probabilidade.

2.2.4 Modelo de Risco Proporcional Dependente do Tempo

Parreira (2007) propôs o modelo de risco proporcional dependente do tempo, sendo que a função de risco é dada por

$$h(t|\alpha,\beta) = \exp(\alpha t + \mathbf{x}'\boldsymbol{\beta}). \tag{2.16}$$

Quando

$$\begin{cases} q(\alpha,\boldsymbol{\beta}) \to 1 \quad \text{e} \\ \lambda = 1, \end{cases}$$

temos que o modelo GTDL se aproxima do modelo de risco proporcional dependente do tempo. Notemos que o modelo de risco proporcional dependente do tempo pode ser um caso particular do modelo de Gompertz, quando $\eta = \alpha$ e $\phi = \mathbf{x}'\boldsymbol{\beta}$.

A função de sobrevivência para o modelo de risco dado em (2.16) é dada por

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

e a função de densidade de probabilidade é dada por

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

Um exemplo das curvas das funções de risco, sobrevivência e de densidade de probabilidade pode ser observado nas Figuras 2.8 e 2.9, respectivamente.

2.2.5 Verificação de Proporcionalidade

Alguns métodos foram propostos na literatura para a verificação de risco proporcional. Apresentamos um método gráfico com essa finalidade. Este método foi retirado de Colosimo e Giolo (2006).

Método Gráfico Descritivo

Este método consiste em construir um gráfico da seguinte forma,

• dividir os dados em m estratos, usualmente de acordo com alguma covariável, por exemplo, dividir os dados em dois estratos de acordo com a covariável sexo;

• estimar $H(t)$ para cada estrato utilizando o estimador de Nelson-Aalen-Breslow (Colosimo e Giolo, 2006);

• construir o gráfico logaritmo de $\widehat{H(t)}$ *versus* t, ou $\log(t)$;

Se existir curvas não paralelas então existe indícios de não proporcionalidade dos riscos. É razoável construir este gráfico para cada covariável incluída no estudo.

2.3 Aplicação

Para a aplicação da teoria desenvolvida neste Capítulo utilizamos o conjunto de dados encontrado no *software* R Development Core Team (2008) com o nome de *lung* (Loprinzi *et al.*, 1994). Os tempos registrados são de 228 pacientes com câncer de pulmão, da Clínica Mayo, sendo que para 63 pacientes os tempos foram censurados.

Após utilizar procedimentos de seleção de covariáveis usando o modelo GTDL, apenas a covariável sexo permaneceu na modelagem.

Para verificar a suposição de risco não proporcional fazemos o uso do método descrito na Seção 2.2.5. O resultado é apresentado na Figura 2.10.

O uso deste método é subjetivo pois, a análise gráfica para diferentes pessoas podem levar a diferentes resultados. Neste caso adotamos que as curvas não são paralelas, assim concluindo a não proporcionalidade.

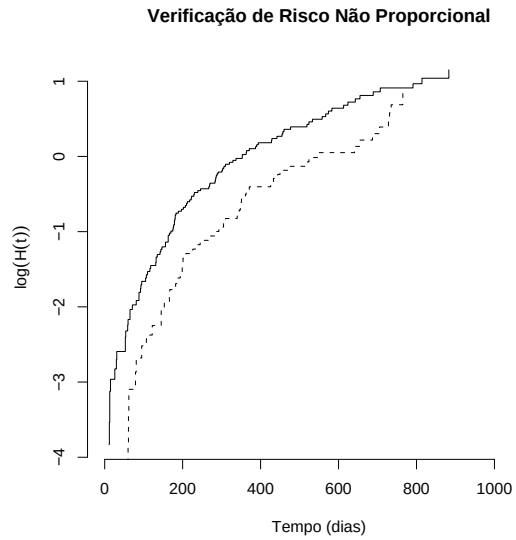

Figura 2.10: Verificação de não proporcionalidade.

Como temos que o parâmetro λ assume valores apenas nos $\mathbb{R}^+$ então utilizamos a transformação $\lambda = \exp(\phi)$, para termos ϕ assumindo valores em $\mathbb{R}$. Esta transformação facilita o processo de otimização da log verossimilhança.

Usamos a inferência clássica para a estimação dos parâmetros, para a construção dos intervalos de confiança assumimos normalidade assintótica. Os resultados como estimador de máxima verossimilhança (EMV), desvio padrão (DP) e intervalo de confiança (IC) são mostrados na Tabela 2.2. A comparação entre as curvas estimadas pelo estimador de Kaplan-Meier e as curvas traçadas utilizando o estimador de máxima verossimilhança é apresentada na Figura 2.11.

Tabela 2.2: Estimativas clássicas da aplicação do modelo GTDL.

Parâmetros	EMV	DP	IC(95%)
ϕ	$-5,5932$	$0,1438$	$[-5,8750;\ -5,1146]$
α	$0,0072$	$0,0035$	$[0,0004;\ 0,01400]$
β	$-1,6898$	$0,4500$	$[-2,5717;\ -0,8080]$

Notamos pela Tabela 2.2 que todos os parâmetros são significativos, como o parâmetro α é positivo o tempo tem o efeito de aceleração para que a falha aconteça. Pela Figura 2.11 vemos que as curvas traçadas utilizando o estimador de máxima verossimilhança se comporta de maneira parecida com a curva estimada pelo estimador de

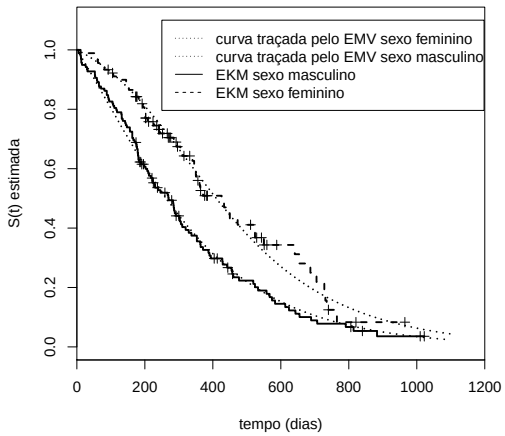

Figura 2.11: Estimativas das funções de sobrevivência para a aplicação.

Kaplan-Meier.

2.4 Considerações Finais

Neste Capítulo apresentamos o modelo GTDL que tem a razão das funções de risco de dois indivíduos não proporcional e estudamos o comportamento das funções de risco e sobrevivência do mesmo. Os modelos de Gompertz, exponencial, taxa de cura e de risco proporcional dependente do tempo foram identificados como casos particulares do modelo GTDL e a partir de um conjunto de dados reais ajustamos o modelo apresentado.

Capítulo 3

Modelo de Risco Logístico Generalizado Dependente do Tempo com Fragilidade

O modelo de risco logístico generalizado dependente do tempo é uma generalização da família função logística denominado em Kalbfleisch e Prentice (1980). Tal modelo foi estudado por Ha e MacKenzie (2010) incluindo a interpretação de fragilidade para dados recorrentes utilizando a verossimilhança hierárquica.

Neste Capítulo apresentamos uma abordagem para o modelo GTDL com fragilidade que pode ser utilizada ou quando se tem apenas uma observação por indivíduo ou quando se tem eventos recorrentes. Na Seção 3.1 é apresentado o modelo de fragilidade multiplicativo. O modelo logístico generalizado dependente do tempo com fragilidade é apresentado na Seção 3.2 e uma aplicação do modelo logístico generalizado dependente do tempo com fragilidade utilizando um conjunto de dados simulado é mostrado na Seção 3.3.

3.1 Modelo de Fragilidade Multiplicativo

Quando estamos trabalhando com modelos de sobrevivência e a suposição que os tempos de falha dos indivíduos são dependentes dada as covariáveis, não devemos utilizar os modelos de Cox (1972) ou Mackenzie (1996), por exemplo, pois estes modelos se baseiam que os indivíduos sejam independentes dada as covariáveis.

Para contornar o problema de dependência Clayton (1978) propôs uma extensão do modelo de Cox (1972) introduzindo um efeito aleatório (fragilidade) de maneira multiplicativa na função de risco, que tem a finalidade de captar a dependência e heterogeneidade não observada. A representação do modelo de Cox (1972) com o termo de fragilidade multiplicativo é dado por

$$h_i(t|v_i, \mathbf{x}_i) = v_i h_0(t) \exp(\mathbf{x}_i'\boldsymbol{\beta}),$$

sendo que v_i representa a fragilidade do *i-ésimo* indivíduo, $h_0(t)$ representa a função de risco de base, $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ é o vetor de coeficientes a serem estimados e $\mathbf{x}_i'$ as covariáveis associadas ao *i-ésimo* indivíduo.

O papel da fragilidade no cenário de tempos univariados é medir uma possível heterogeneidade de modo a identificar a influência das covariáveis que ou não foram incorporadas na modelagem ou não podem ser medidas.

Como a variável de fragilidade, representada neste trabalho por v_i, é uma variável latente, precisamos adotar uma distribuição conhecida a mesma, várias distribuições foram propostas como, Hougaard (1984) sugeriu a distribuição gaussiana inversa , Korsgaard *et al.* (1998) utilizou a distribuição log-normal, Hougaard (1986a, b) sugeriu uma distribuição que generaliza, como por exemplo, a gama e a gaussiana inversa e Aalen (1988), propôs uma distribuição que tem como caso particular a distribuição sugerida por Hougaard (1986a, b). Porém, a mais utilizada é a distribuição gama, sendo que Clayton (1978), Clayton e Cuzick (1985) e Tomazella *et al.* (2004) são alguns dos autores que utilizaram tal distribuição.

Um inconveniente quando se trabalha com fragilidade é a obtenção da função de risco não condicional a v_i, mas sim ao parâmetro da distribuição de fragilidade adotada. Uma saída é a utilização da transformada de Laplace, pois esta tem a mesma forma que a função de sobrevivência não condicional, mas a obtenção da função de sobrevivência não condicional de maneira explicita não ocorre para todas as distribuições de fragilidade. Algumas outras maneiras de se trabalhar com modelos de fragilidade podem ser vista em Tomazella (2003) e Ha *et al.* (2001).

Esta maneira de incorporar a fragilidade na função de risco tem sido usada na maioria dos trabalhos em análise de sobrevivência quando existe dependência no conjunto de dados. Quando $v_i > 1$, temos que o indivíduo i é mais frágil e, se torna mais resistente

quando $v_i < 1$, daí o nome de modelo de fragilidade, pois quanto maior o v_i mais frágil se torna a unidade i.

3.2 Modelo GTDL com Fragilidade Gama

Os modelos com fragilidade são construídos a partir da inclusão de um efeito aleatório na função de risco, este é considerado ser não negativo. Assumimos que o efeito aleatório segue alguma distribuição, sendo que a mais utilizada na literatura é a distribuição gama. Neste trabalho utilizamos a distribuição gama pois, no ponto de vista computacional, ela se ajusta muito bem para modelos de sobrevivência, por causa da sua conveniência algébrica. Isso é devido à simplicidade das derivadas da transformada de Laplace.

Considere que V tenha distribuição gama com parâmetro τ e η independentes com $\tau \geq 0$ e $\eta \geq 0$. A função de densidade de probabilidade é dada por,

$$f(v) = \frac{\eta^{\tau}}{\Gamma(\tau)} v^{(\tau-1)} \exp(-v\eta). \tag{3.1}$$

Para satisfazer a suposição de identificabilidade para modelos com fragilidade proposta por Elbers e Ridder (1982), precisamos que $\mathrm{E}(V) = 1$. Adotando $\tau = \eta = \theta^{-1}$, a função de densidade de probabilidade (3.1) é dada por,

$$f(v) = \frac{\left(\frac{1}{\theta}\right)^{1/\theta}}{\Gamma\left(\frac{1}{\theta}\right)} v^{(\frac{1}{\theta}-1)} \exp(-v/\theta), \tag{3.2}$$

e assim $\mathrm{E}(V)$=1 e $\mathrm{Var}(V)$=θ. A variância da distribuição gama é o parâmetro de dependência ou de heterogeneidade não observada. Observe que, se $\theta = 0$, temos que todas as variáveis de fragilidade são iguais a 1, ou seja, a distribuição gama fica degenerada no ponto 1 e com isso, temos o modelo sem fragilidade.

A partir do modelo GTDL dado pela equação (2.1) temos que a função de risco do *i-ésimo* indivíduo com o termo de fragilidade multiplicativo v_i, é dado por

$$h_i(t|\alpha, \boldsymbol{\beta}, \lambda, v_i) = v_i \frac{\lambda \exp(\alpha t + \mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\alpha t + \mathbf{x}_i'\boldsymbol{\beta})}, \tag{3.3}$$

Este risco individual é interpretado como a função de risco condicional do *i-ésimo* indivíduo dado v_i. A correspondente função de sobrevivência condicional é dada por

$$S_i(t) = S_i(t|v_i) = S_0(t)^{v_i},$$

ou seja,

$$S_i(t|\alpha, \boldsymbol{\beta}, \lambda, v_i) = \left(\frac{1 + \exp(\alpha t + \mathbf{x}_i' \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})} \right)^{\frac{-\lambda v_i}{\alpha}}, \tag{3.4}$$

representando a probabilidade do indivíduo i sobreviver até o tempo t dado o efeito aleatório v_i.

Para obtermos a função de sobrevivência não condicional precisamos integrar o termo de fragilidade, da forma

$$S(t) = \int_0^\infty S(t|v) g(v) dv, \tag{3.5}$$

sendo que $g(v)$ é a função de densidade de probabilidade da gama dada pela equação (3.2).

Para que possamos resolver a equação (3.5) utilizamos a transformada de Laplace, ferramenta muito útil para a modelos com fragilidade.

Dada uma função $f(x)$, a transformada de Laplace considera a função de um argumento real s e é definida como

$$Q(s) = \int_0^\infty e^{-sx} f(x) dx. \tag{3.6}$$

A razão da transformada de Laplace ser muito útil nesta situação é porque apresenta a mesma forma que a função de sobrevivência não condicional.

Na equação (3.6), suponha que $f(x)$ seja a densidade da fragilidade V, ou seja, $f(x) = g(v)$ e s é o risco acumulado, $H(t)$. Com isso obtemos,

$$S(t) = \int_0^\infty e^{-H(t)v} f(v) dv = Q(H(t)).$$

A transformada de Laplace da densidade Gama$(1/\theta, 1/\theta)$, é dada por

$$\begin{aligned} Q(s) &= \left(\frac{\frac{1}{\theta}}{\frac{1}{\theta} + s} \right)^{\frac{1}{\theta}} \\ &= (1 + \theta s)^{-1/\theta}, \end{aligned} \tag{3.7}$$

Substituindo $s = H(t)$ na equação (3.7), e após alguns cálculos obtemos a função de sobrevivência não condicional, dada por,

$$S(t|\alpha, \boldsymbol{\beta}, \lambda, \theta) = \left[1 + \frac{\lambda \theta}{\alpha} \log \left(\frac{1 + \exp(\alpha t + \mathbf{x}' \boldsymbol{\beta})}{1 + \exp(\mathbf{x}' \boldsymbol{\beta})} \right) \right]^{-\frac{1}{\theta}}, \tag{3.8}$$

e utilizando a relação (1.5), a função de risco é dada por,

$$h(t|\alpha,\boldsymbol{\beta},\lambda,\theta)=\frac{\lambda\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{\left[1+\frac{\lambda\theta}{\alpha}\log\left(\frac{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)\right](1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta}))}. \tag{3.9}$$

Usando a relação (1.5) a função de densidade de probabilidade é expressa por

$$f(t|\alpha,\boldsymbol{\beta},\lambda,\theta)=\frac{\lambda\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{(1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta}))}\left[1+\frac{\lambda\theta}{\alpha}\log\left(\frac{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)\right]^{-(1/\theta+1)}. \tag{3.10}$$

Notamos na equação (3.8) que quando $\alpha<0$ o modelo GTDL se torna uma modelo de longa duração, e a taxa de cura, p, é da forma

$$\begin{aligned} p=\lim_{t\longrightarrow\infty}S(t|\mathbf{x},\alpha,\boldsymbol{\beta},\lambda,\theta) &= \lim_{t\longrightarrow\infty}\left[1+\frac{\lambda\theta}{\alpha}\log\left(\frac{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)\right]^{-\frac{1}{\theta}} \\ &= \left[1+\frac{\lambda\theta}{\alpha}\ln\left(\frac{1}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)\right]^{\frac{-1}{\alpha}}. \end{aligned} \tag{3.11}$$

A Figura 3.1 (a) representa algumas formas da função de sobrevivência dada em (3.8), a linha preta contínua significa a taxa de curados, utilizando a equação (3.11) a taxa de curados é igual a 0,30. As formas da função de risco dada em (3.9) estão representadas na Figura 3.1 (b). Já vimos na Seção 2.2 que a função de risco do modelo GTDL apresenta as formas crescente e constante. Com a introdução da fragilidade, a função de risco do modelo resultante passa a ter o comportamento unimodal.

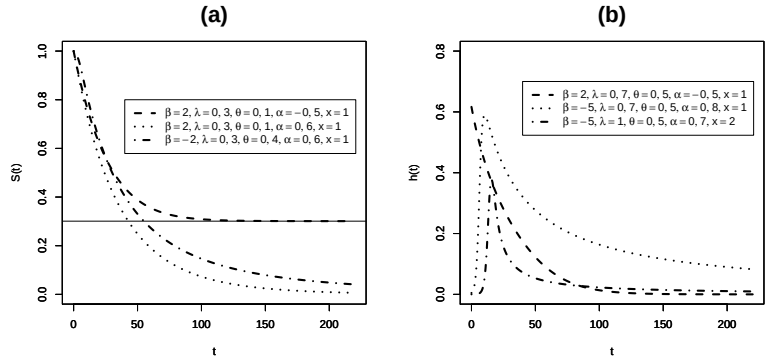

Figura 3.1: Funções de sobrevivência em (a) e de risco em (b) do modelo GTDL com fragilidade.

As formas da função de densidade de probabilidade dada em (3.10) estão apresentadas na Figura 3.2.

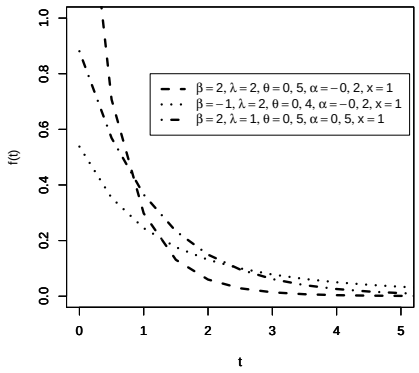

Figura 3.2: Formas da função de densidade de probabilidade do modelo GTDL com fragilidade.

Utilizando as funções de sobrevivência e de risco não condicional, a função de verossimilhança é dada por

$$\begin{aligned} L(\alpha,\boldsymbol{\beta},\lambda,\theta|\text{dados}) &= \prod_{i=1}^{n}[h(t_i;\lambda,\alpha,\boldsymbol{\beta},\theta)]^{\delta_i}[S(t_i;\lambda,\alpha,\boldsymbol{\beta},\theta)] \\ &= \prod_{i=1}^{n}\left[\frac{\lambda\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}\right]^{\delta_i}\left[1+\frac{\lambda\theta}{\alpha}\log\left(\frac{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)\right]^{-(1/\theta+\delta_i)} \end{aligned} \quad (3.12)$$

Denotando $l(\alpha,\boldsymbol{\beta},\lambda,\theta|\text{dados}) = log(L(\alpha,\boldsymbol{\beta},\lambda,\theta|\text{dados}))$, e após alguns cálculos, obtemos que

$$\begin{aligned} l(\alpha,\boldsymbol{\beta},\lambda,\theta|\text{dados}) &= log(\lambda)\sum_{i=1}^{n}\delta_i+\sum_{i=1}^{n}\delta_i(\mathbf{x}'\boldsymbol{\beta}+\alpha t_i)-\sum_{i=1}^{n}\delta_i\log(1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})) \\ &- \sum_{i=1}^{n}(\delta_i+1/\theta)\left\{log\left[1+\frac{\theta\lambda}{\alpha}\log\left(\frac{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)\right]\right\}. \end{aligned} \quad (3.13)$$

Para encontrar estimativas para os parâmetros da função (3.13) usamos métodos iterativos como por exemplo, o de Newton-Raphson. Este método já está implementado em alguns *softwares*. Para usar este método precisamos das primeiras derivadas parciais. Utilizando as relações (2.5), (2.6), (2.7) e denotando-as por p, q e g, respectivamente, as derivadas parciais da função (3.13) com relação aos parâmetros são:

$$\frac{\partial l}{\partial\alpha} = \sum_{i=1}^{n}(\delta_i t_i)-\sum_{i=1}^{n}(\delta_i p t_i)-\sum_{i=1}^{n}\left[\frac{\frac{-\lambda\theta}{\alpha^2}\ln\left(\frac{1}{qg}\right)+\frac{\lambda\theta}{\alpha}t_i p}{1+\frac{\lambda\theta}{\alpha}\ln\left(\frac{1}{qg}\right)}(1/\theta+\delta_i)\right]; \quad (3.14)$$

$$\begin{aligned}\frac{\partial l}{\partial \beta} &= \sum_{i=1}^{n}(\delta_i \mathbf{x}_i) + \sum_{i=1}^{n}\left(\frac{\delta_i \mathbf{x}_i \exp(\mathbf{x}_i'\boldsymbol{\beta})}{g}\right) \\ &- \sum_{i=1}^{n}\left\{\frac{\frac{\lambda\theta}{\alpha}\left[\frac{-\mathbf{x}_i \exp(\mathbf{x}_i'\boldsymbol{\beta})}{q} + \mathbf{x}_i \exp(\alpha t_i + \mathbf{x}_i'\boldsymbol{\beta})g\right]}{1 + \frac{\lambda\theta}{\alpha}\ln\left(\frac{1}{qg}\right)g}(\delta_i + 1/\theta)\right\}; \end{aligned} \tag{3.15}$$

$$\frac{\partial l}{\partial \theta} = -\sum_{i=1}^{n}\left\{\frac{\frac{\lambda}{\alpha}\ln\left(\frac{1}{qg}\right)}{1 + \frac{\lambda\theta}{\alpha}\ln\left(\frac{1}{qg}\right)}(1/\theta + \delta_i)\right\} + (1/\theta^2)\sum_{i=1}^{n}\ln\left[1 + \frac{\lambda\theta}{\alpha}\ln\left(\frac{1}{qg}\right)\right]; \tag{3.16}$$

$$\frac{\partial l}{\partial \lambda} = -\sum_{i=1}^{n}\left\{\frac{\frac{\theta}{\alpha}\ln\left(\frac{1}{qg}\right)}{1 + \frac{\lambda\theta}{\alpha}\ln\left(\frac{1}{qg}\right)}(1/\theta + \delta_i)\right\} + \frac{1}{\lambda}\sum_{i=1}^{n}(\delta_i). \tag{3.17}$$

Para encontrar as estimativas de máxima verossimilhança precisamos igualar as derivadas parciais iguais a zero e resolver o sistema. Devido à complexidade das expressões (3.14) a (3.17) métodos iterativos são necessários para resolver o sistema. Para encontrar as estimativas intervalares utilizamos a matrix de informação observada pois a matriz de informação de Fisher não é facilmente obtida por isso.

3.3 Aplicação com Dados Simulados

Apenas para efeito de exemplificação da teoria desenvolvida neste Capítulo utilizamos um conjunto de dados gerados do modelo GTDL com fragilidade. Para a geração fixamos

$$\alpha = 0,10;\ \beta = -3,00;\ \lambda = 0,50 \text{ e } \theta = 0,50;$$

sendo que o tamanho da amostra é 150 e a covariável foi gerada da distribuição Bernoulli com probabilidade de sucesso igual a 0,60. Utilizamos o método de geração de dados do modelo apresentado na Seção 4.1.1.

Para verificação de risco proporcional usamos o método descrito na Seção 2.2.5 e o resultado é apresentado na Figura 3.3. Notamos que as curvas não são paralelas e

com isso temos que os dados suportam a suposição de risco não proporcional, como já era esperado, pois geramos os dados do modelo GTDL com fragilidade que é um modelo de risco não proporcional.

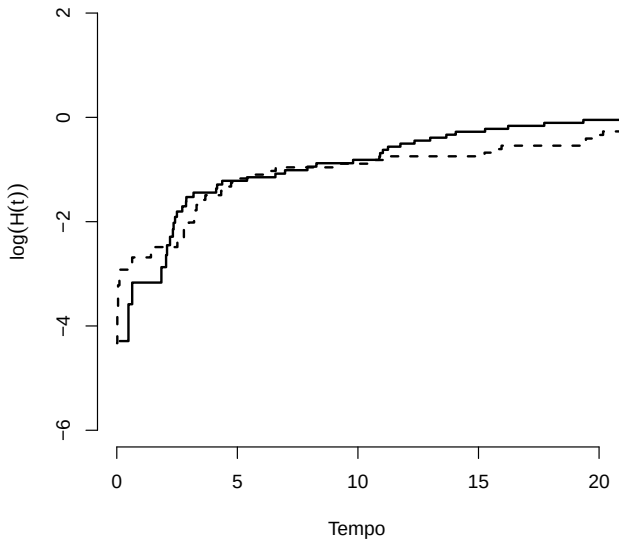

Figura 3.3: Verificação de riscos não proporcional.

Utilizamos a transformação $\theta = \exp(\theta^*)$ e $\lambda = \exp(\lambda^*)$ para que os parâmetros θ^* e λ^* sejam definidos nos **R**. Permitindo-nos interpretações, como por exemplo, se o parâmetro é significativo ou não. Esta transformação é utilizada em todos os estudos sob a abordagem clássica.

Usando inferência clássica e o *software* SAS Institute Inc. obtivemos o estimador de máxima verossimilhança (EMV), desvio padrão (DP) e o intervalo de confiança (IC) de 95% para cada parâmetro, sendo apresentado na Tabela 3.1.

Tabela 3.1: Resultados do ajuste do modelo com dados simulados.

Parâmetros	EMV	DP	IC(95%)
θ^*	$-0,5253$	$0,2014$	$[-0,9200;\ -0,1306]$
α	$0,1037$	$0,0253$	$[0,0285;\ 0,1595]$
β	$-3,1074$	$0,4582$	$[-4,0054;\ -2,2093]$
λ^*	$-1,2216$	$0,2481$	$[-1,7080;\ -0.7353]$

Observamos pela Tabela 3.1 que os EMV estão próximos do valor verdadeiro e os intervalos de confiança cobrem os verdadeiros valores dos parâmetros.

3.4 Considerações Finais

Apresentamos neste Capítulo o modelo GTDL com fragilidade multiplicativa quando temos apenas uma observação por indivíduo. Para ilustrar este modelo foi feita uma aplicação com dados gerados com o intuito de verificar se as estimativas dos parâmetros ficariam perto do verdadeiro valor. Os resultados obtidos foram satisfatórios.

Capítulo 4

Aplicações

Desenvolvemos na Seção 4.1 um estudo de simulação onde, um método de geração de dados do modelo GTDL com fragilidade é apresentado em 4.1.1, o cálculo da probabilidade de cobertura, o vício e o erro quadrático médio se encontra em 4.1.2. Ajustamos os modelos GTDL e GTDL com fragilidade e fizemos uma comparação entre eles utilizando um conjunto de dados reais na Seções 4.2.

4.1 Um Estudo de Simulação

4.1.1 Geração de Dados do Modelo GTDL com Fragilidade

Vários métodos de geração de dados são propostos na literatura. Gamerman (1996) apresenta o método da rejeição, o método da rejeição adaptativo, entre outros.

Neste estudo a geração de dados do modelo GTDL com fragilidade é feito utilizando o método da função inversa, ou seja, se $u^* \sim U(0,1)$ então

$$u^* = F(t) = 1 - S(t). \tag{4.1}$$

O resultado (4.1) só é válido quando a função de sobrevivência é própria, ou seja, para o modelo em estudo a restrição é satisfeita quando $\alpha > 0$.

Utilizando $u = 1 - u^* \sim U(0,1)$, então

$$\begin{aligned} u &= \left(\frac{1+\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)^{\frac{-\lambda v}{\alpha}} \\ \Rightarrow t &= \frac{1}{\alpha}\left\{-\mathbf{x}'\boldsymbol{\beta} + \log\left[-1 + \left((1+\exp(\mathbf{x}'\boldsymbol{\beta}))\exp\left(\frac{-\alpha\log(u)}{\lambda v}\right)\right)\right]\right\}. \end{aligned} \tag{4.2}$$

Para encontrarmos o tempo, t, precisamos fixar valores para λ, α, $\boldsymbol{\beta}$, gerar um valor v da distribuição gama($1/\theta$,$1/\theta$), gerar um valor u da distribuição $U(0,1)$, e ainda é necessário gerar o valor da covariável $\mathbf{x}$. A covariável pode ser gerada da distribuição uniforme ou da normal por exemplo. A substituição destas quantidades em (4.2), resulta na obtenção de t.

Este método de geração não leva em consideração a censura, ou seja, todos os tempos gerados pelo método são de falha. Com isso a variável indicadora de censura sempre será igual a 1. Quando o interesse é um conjunto de dados cujos tempos são ou de falha ou de censura então a geração dos tempos pode ser feita da seguinte forma:

1. gerar um tempo t_1 utilizando a equação 4.2;
2. gerar um tempo t_2 pela distribuição exponencial;
3. encontar $t = \min\{t_1, t_2\}$;
4. adotar a variável indicadora de censura que assumirá o seguinte valor

$$\delta = \begin{cases} 1, & \text{se } t = t_1 \\ 0, & \text{se } t = t_2. \end{cases}$$

4.1.2 Cálculo da Probabilidade de Cobertura

Para a determinação da probabilidade de cobertura dos parâmetros foram geradas 1.000 amostras com diferentes tamanhos. Para cada amostra estimamos os parâmetros, e os intervalos de confiança foram construídos assumindo que os estimadores têm normalidade assintótica. Verificamos se o valor do verdadeiro parâmetro está contido no intervalo de confiança e a quantidade de vezes que isso ocorre é calculada. Com isso, a probabilidade de cobertura é encontrada da seguinte forma, como por exemplo para o parâmetro α,

$$pc_\alpha = \frac{\text{nº de intervalos que contém o verdadeiro valor do parâmetro } \alpha}{\text{nº total de intervalos construídos}},$$

sendo que pc_α significa probabilidade de cobertura do parâmetro α. Os intervalos foram construídos com 95% de confiança.

Para a geração fixamos

$$\alpha = 0,10;\ \beta = -3,00;\ \lambda = 0,50 \text{ e } \theta = 0,50;$$

e geramos a covariável a partir da distribuição Bernoulli com probabilidade de sucesso igual a 0,5.

Neste estudo consideramos dados com censura (10% e 30%) e dados sem censura, a probabilidade de cobertura para cada tamanho de amostra e parâmetro são mostrados nas Tabelas 4.1 a 4.3.

Tabela 4.1: Probabilidade de cobertura com dados sem censura.

	$n = 50$	$n = 100$	$n = 300$	$n = 500$
pc_α	0,8670	0,9120	0,9490	0,9490
pc_β	0,9380	0,9430	0,9550	0,9610
pc_{θ^*}	0,9680	0,9730	0,9740	0,9780
pc_{λ^*}	0,9130	0,9400	0,9580	0,9590

Tabela 4.2: Probabilidade de cobertura com dados com 10% de censura.

	$n = 50$	$n = 100$	$n = 300$	$n = 500$
pc_α	0,8810	0,8960	0,9430	0,9420
pc_β	0,9230	0,9370	0,9440	0,9590
pc_{θ^*}	0,9540	0,9610	0,9650	0,9740
pc_{λ^*}	0,9060	0,9290	0,9510	0,9570

Tabela 4.3: Probabilidade de cobertura com dados com 30% de censura.

	$n = 50$	$n = 100$	$n = 300$	$n = 500$
pc_α	0,9030	0,9130	0,9450	0,9500
pc_β	0,9380	0,9400	0,9450	0,9600
pc_{θ^*}	0,9430	0,9590	0,9670	0,9750
pc_{λ^*}	0,9070	0,9400	0,9530	0,9580

Para cada tamanho de amostra as representações gráficas para as probabilidades de cobertura dos parâmetros, são mostradas nas Figuras 4.1 e 4.2. Nestas Figuras as duas linhas pretas contínuas representam os limites inferior e superior de 1.000 amostras da distribuição Bernoulli com probabilidade de sucesso igual a 0,95.

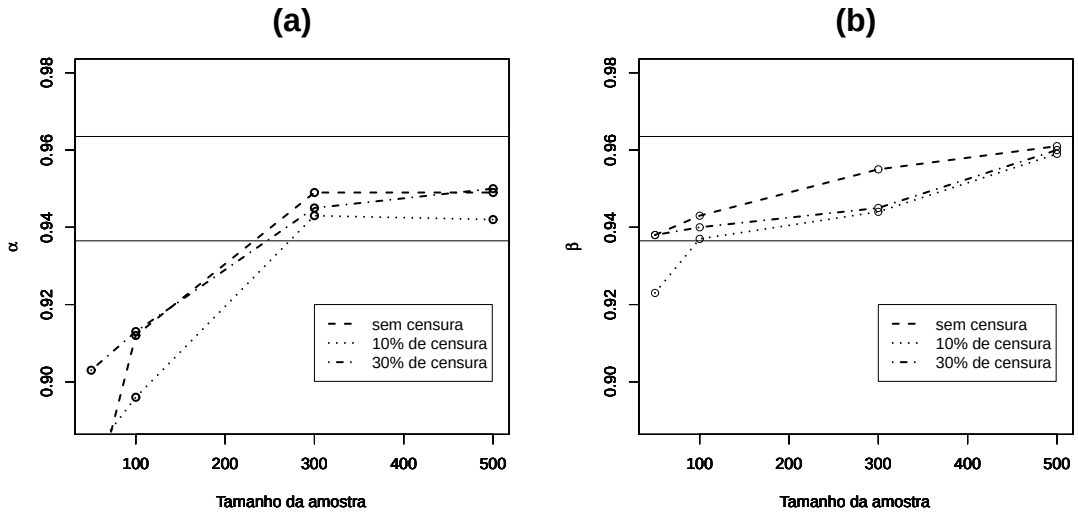

Figura 4.1: Probabilidade de cobertura de α em (a) e de β em (b).

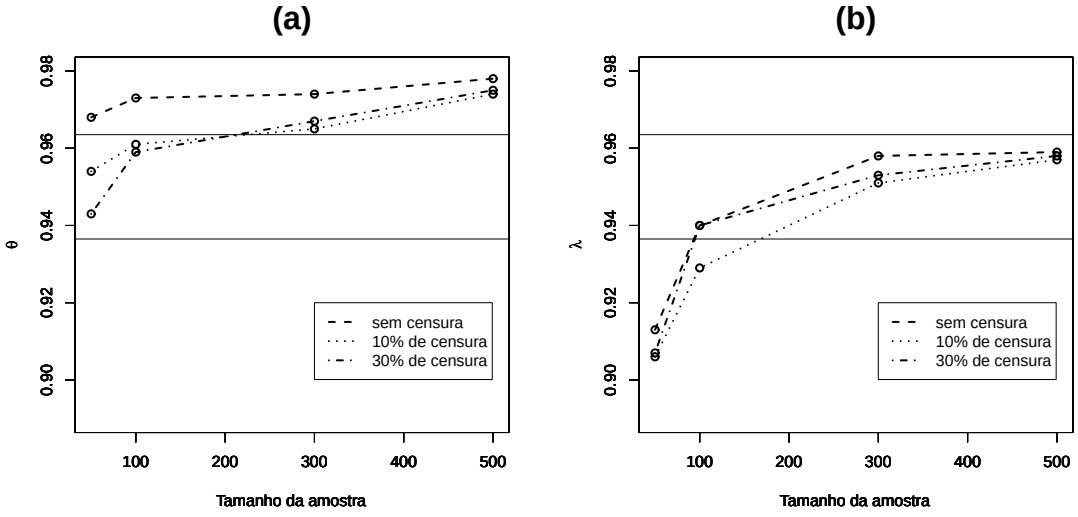

Figura 4.2: Probabilidade de cobertura de θ em (a) e de λ em (b).

Observamos que o comportamento da probabilidade de cobertura não se modifica quando comparamos os dados censurados dos dados não censurados. Para tamanhos de amostra com 300 ou mais observações, a probabilidade de cobertura calculada fica dentro ou acima dos limites da probabilidade nominal.

Utilizando a mesma rotina de programação também calculamos o vício das estimativas dos parâmetros, nas três situações (sem censura, com 10% e 30% de dados censurados) cujos resultados se encontram nas Tabelas 4.4 a 4.6.

Tabela 4.4: Vício com dados sem censura.

Parâmetros	$n = 50$	$n = 100$	$n = 300$	$n = 500$
α	-0,0071	-0,0058	-0,0012	-0,0004
β	0,0484	0,0585	0,0133	-0,0017
θ^*	1,5703	0,3965	0,0642	0,0273
λ^*	0,0512	0,0286	0,0040	0,0061

Tabela 4.5: Vício com dados com 10% de censura.

Parâmetros	$n = 50$	$n = 100$	$n = 300$	$n = 500$
α	-0,0156	-0,0036	-0,0011	-0,0005
β	0,0550	0,0316	0,0026	0,0050
θ^*	1,7312	0,6503	0,0798	0,0528
λ^*	0,0662	0,0289	0,0153	0,0075

Tabela 4.6: Vício com dados com 30% de censura.

Parâmetros	$n = 50$	$n = 100$	$n = 300$	$n = 500$
α	-0,0472	-0,0083	-0,0025	$< 0,0001$
β	0,3386	0,0500	0,0199	0,0008
θ^*	2,2180	0,8502	0,1094	0,0712
λ^*	0,0457	0,0351	0,0063	0,0060

Verificamos que o vício diminui tendendo para zero quando aumentamos o tamanho da amostra; verificamos isso para as três situações e para todos os parâmetros.

Uma forma de verificar o quão próximo as estimativas dos parâmetros estão do verdadeiro valor é calcular o erro quadrático médio (EQM). A formulação do erro quadrático médio, como por exemplo, para o parâmetro β é,

$$EQM = \sum_{n=1}^{N} \frac{(\widehat{\beta} - \beta)^2}{N - 1},$$

sendo que N é a quantidade de repetições. No nosso caso $N = 1.000$.

Ainda com a mesma rotina calculamos o erro quadrático médio e os resultados se encontram nas Tabelas 4.7 a 4.9.

Tabela 4.7: Erro quadrático médio (EQM) com dados sem censura.

Parâmetros	$n = 50$	$n = 100$	$n = 300$	$n = 500$
α	0,0048	0,0016	0,0004	0,0003
β	0,4953	0,2328	0,0701	0,0460
θ^*	13,4430	2,3499	0,0643	0,0320
λ^*	0,1289	0,0609	0,0177	0,0117

Tabela 4.8: Erro quadrático médio (EQM) com dados com 10% de censura.

Parâmetros	$n = 50$	$n = 100$	$n = 300$	$n = 500$
α	0,0117	0,0022	0,0005	0,0003
β	0,6251	0,2610	0,0788	0,0456
θ^*	15,8823	4,4143	0,0896	0,0498
λ^*	0,1329	0,0658	0,0185	0,0109

Tabela 4.9: Erro quadrático médio (EQM) com dados com 30% de censura.

Parâmetros	$n = 50$	$n = 100$	$n = 300$	$n = 500$
α	0,4013	0,0032	0,0008	0,0004
β	1,5236	0,3077	0,1010	0,0530
θ^*	21,4399	6,4997	0,2445	0,0841
λ^*	0,1427	0,0701	0,0228	0,0137

Notamos pelos resultados que quando o tamanho da amostra cresce o EQM diminua e, o vício diminui tendendo para zero; com isso notamos que a parcela do vício no EQM diminui, mostrando que o estimador é não viciado. Isso ocorre para todos os parâmetros.

Também utilizando a mesma rotina de programação para calcular o tamanho do intervalo de confiança. Os resultados são apresentados nas Tabelas 4.10 a 4.12.

Tabela 4.10: Tamanho do intervalo de confiança com dados sem censura.

Parâmetros	$n = 50$	$n = 100$	$n = 300$	$n = 500$
α	0,1967	0,1388	0,0772	0,0594
β	2,6220	1,8565	1,0390	0,8011
θ^*	17,1536	3,8745	0,8992	0,6626
λ^*	1,3015	0,9431	0,5378	0,4177

Tabela 4.11: Tamanho do intervalo de confiança com dados com 10% de censura.

Parâmetros	$n = 50$	$n = 100$	$n = 300$	$n = 500$
α	0,2378	0,1492	0,0843	0,0649
β	2,7819	1,9006	1,0810	0,8332
θ^*	21,3105	6,8134	1,0541	0,7805
λ^*	1,3190	0,9481	0,5493	0,4249

Tabela 4.12: Tamanho do intervalo de confiança com dados com 30% de censura.

Parâmetros	$n = 50$	$n = 100$	$n = 300$	$n = 500$
α	0,2856	0,1877	0,1065	0,0810
β	3,0923	2,1012	1,1927	0,9143
θ^*	24,9508	10,3290	1,5394	1,0497
λ^*	1,4098	1,0120	0,5816	0,4493

Observamos pelas Tabelas 4.10 a 4.12 que com o aumento do tamanho da amostra o tamanho do intervalo de confiança diminui, isso ocorre para todos os parâmetros nos três cenários (sem censura, com 10% e 30% de censura). As Figuras 4.3 e 4.4 foram construídas para facilitar a comparação entre os tamanhos dos intervalos de confiança em relação aos dados que contém ou não censura.

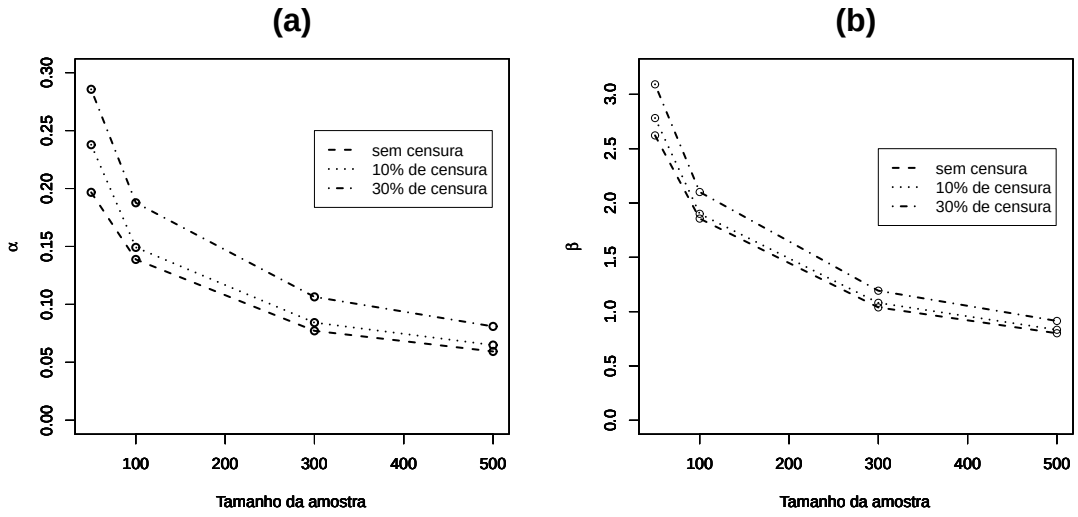

Figura 4.3: Tamanho do intervalo de confiança de α em (a) e de β em (b).

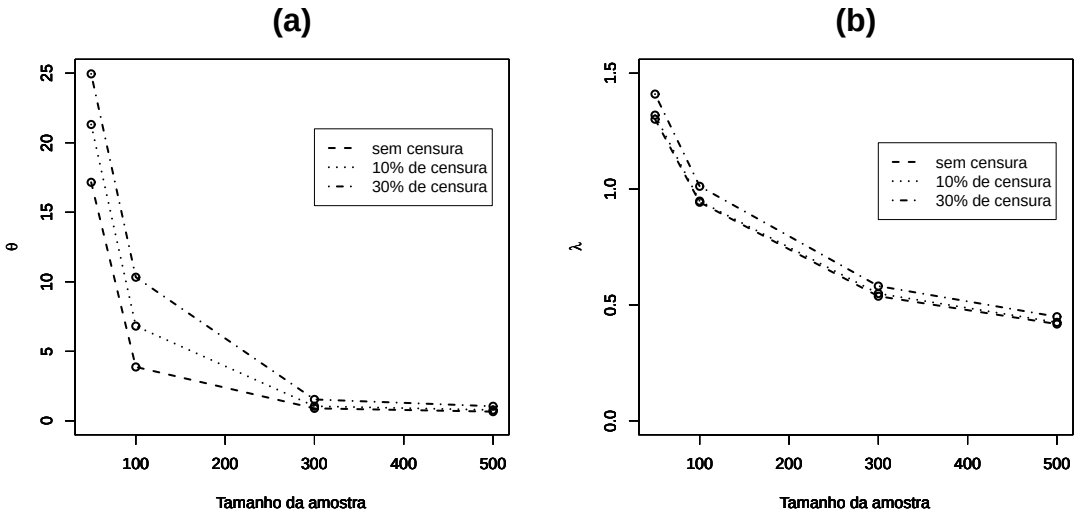

Figura 4.4: Tamanho do intervalo de confiança de θ em (a) e de λ em (b).

Notamos pelas Figuras 4.3 e 4.4 que a ordenação dos intervalos de confiança de menor para maior amplitude são os intervalos calculados dos dados que não contém censura, dos dados com 10% e, por último, os dados com 30% de censura. Isso já erra esperado pois, quanto mais censura tivermos nos dados, menor é quantidade de informação fornecida por eles.

4.2 Aplicação em Dados Reais

Para ilustrar o modelo proposto em (3.9), analisamos um conjunto de dados que faz parte do estudo de incidência de câncer de pulmão na Irlanda do Norte, realizado entre 01/10/1991 e 30/09/1992 (Wilkinson, 1995). Foram identificados 900 casos, sendo que 20 casos foram diagnosticados depois da morte do paciente, 25 casos não puderam ser determinados e 104 pacientes tiveram informações perdidas. O total de pacientes analisado foi de 751 (83%). O tempo foi contado em meses até a morte ou a censura de cada paciente. Para cada paciente foram coletados as informações contidas na Tabela A.1, que se encontra no Apêndice A.

O conjunto de dados é composto de 9 covariáveis mas após investigações preliminares, passamos a trabalhar com 7, pois as covariáveis sexo e idade foram retiradas após utilizarmos o método de seleção de covariáveis. O método também apontou que a covariável fumante não é significativa.

Levando em conta a seguinte frase, o tabagismo é o principal fator de risco do câncer pulmonar, sendo responsável por 90% dos casos, incluímos tal covariável no modelo. Outros fatores relacionados são certos agentes químicos, fatores dietéticos, doença pulmonar obstrutiva crônica, fatores genéticos e história familiar de câncer de pulmão (INCA, 2010). Todos estes fatores não foram medidos, o que nos leva a utilizar o modelo GTDL com fragilidade para captar a influência destes fatores.

Antes de ajustar o modelo precisamos verificar se a suposição de riscos não proporcionais se sustenta, para isso utilizamos o método descrito na Seção 2.2.5. O resultado da aplicação do método são mostrados nas Figuras (4.5) a (4.7).

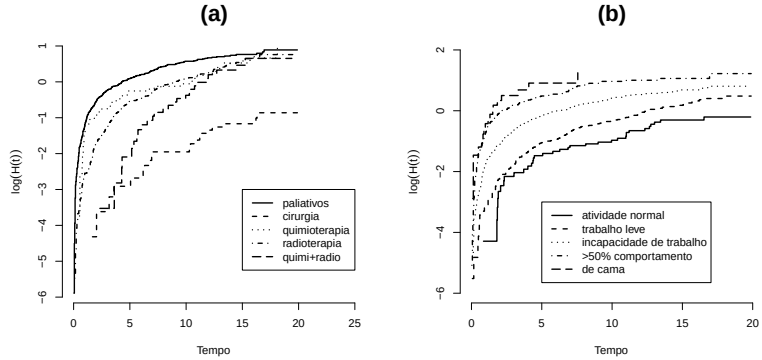

Figura 4.5: Verificação de proporcionalidade para as covariáveis tratamento em (a) e performance em (b).

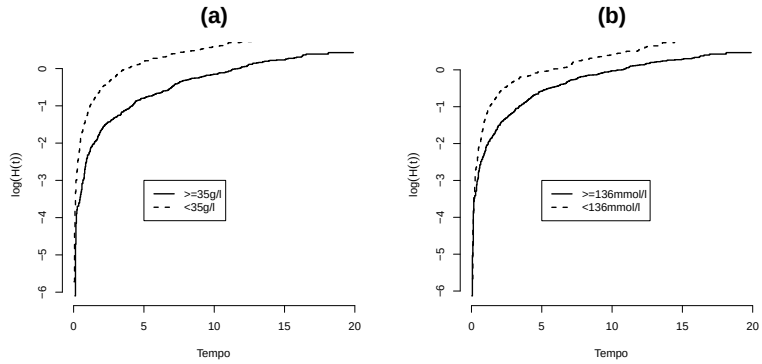

Figura 4.6: Verificação de proporcionalidade para as covariáveis tipo de célula em (a) e nível de sódio em (b).

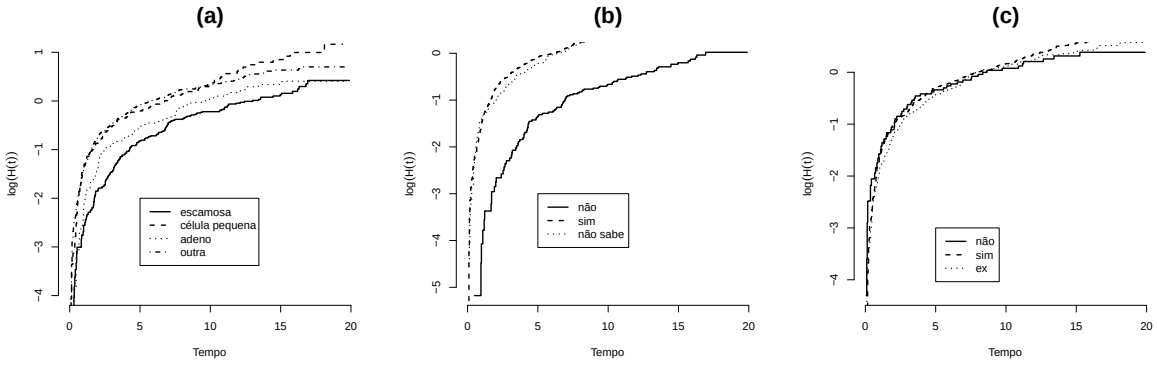

Figura 4.7: Verificação de proporcionalidade para as covariáveis nível de albumina em (a), metástases em (b) e fumante em (c).

Notamos pela Figuras 4.5 (b) e 4.6, que as curvas das covariáveis não se cruzam,

deixando a interpretação de paralelismo um pouco subjetiva, mas em ambas as covariáveis, no inicio do tempo as curvas estão sobrepostas e conforme o tempo aumenta elas vão se distanciando, classificamos este comportamento de não paralelismo. Para as outras covariáveis algumas das curvas se cruzam indicando claramente que os riscos não são proporcionais, isto pode ser visto nas Figuras 4.5 (a) e 4.7, com isso temos que todas as covariáveis indicam risco não proporcional.

A suposição inicial do modelo é satisfeita com isso o modelo GTDL com fragilidade pode ser utilizado. Para ajustar o modelo dado pela equação (3.9) precisamos inserir variáveis *dummys* para representar as categorias das covariáveis, para isso suponha que

$$\beta_2, \beta_3, \ldots, \beta_{19} = 0 \text{ ou } 1,$$

e que as categorias das covariáveis são representadas como se apresenta na Tabela A.2, que se encontra no Apêndice A.

Utilizando inferência clássica e o *software* SAS Institute Inc. obtivemos a estimativa de máxima verossimilhança (EMV), desvio padrão (DP) e o intervalo de confiança (I.C.) de 95% para cada parâmetro, usando os modelos GTDL e GTDL com fragilidade os resultados estão apresentados nas Tabelas 4.14 e 4.13, respectivamente.

Tabela 4.13: Resultado do ajuste do modelo GTDL com fragilidade.

Parâmetros	EMV	DP	IC(95 %)
θ^*	$-0,5151$	$0,2146$	$[-0,9358;\ -0,0944]$
α	$0,1390$	$0,0290$	$[0,0821;\ 0,1958]$
β_3	$1,7185$	$0,4518$	$[0,8330;\ 2,6040]$
β_4	$-0,1450$	$0,4987$	$[-1,1224;\ 0,8323]$
β_5	$0,7559$	$0,4279$	$[-0,0828;\ 1,5947]$
β_6	$1,1846$	$0,4409$	$[0,3205;\ 2,0488]$
β_7	$-4,1188$	$0,5658$	$[-5,2278;\ -3,0099]$
β_8	$-3,9793$	$0,5369$	$[-5,0317;\ -2,9269]$
β_9	$-3,2070$	$0,5173$	$[-4,2209;\ -2,1930]$
β_{10}	$-2,1169$	$0,5712$	$[-3,2364;\ -0,9974]$
β_{11}	$-0,1803$	$0,1865$	$[-0,5459;\ 0,1853]$
β_{12}	$1,0942$	$0,3447$	$[0,4186;\ 1,7698]$
β_{13}	$0,0236$	$0,2260$	$[-0,4193;\ 0,4666]$
β_{14}	$-0,4625$	$0,1539$	$[-0,7641;\ -0,1609]$
β_{15}	$-0,7669$	$0,1667$	$[-1,0936;\ -0,4402]$
β_{16}	$-0,3107$	$0,2305$	$[-0,76260;\ 0,1411]$
β_{17}	$0,6382$	$0,1908$	$[0,2642;\ 1,0122]$
β_{18}	$-0,1374$	$0,2661$	$[-0,6589;\ 0,3841]$
β_{19}	$0,0790$	$0,1540$	$[-0,2228;\ 0,3808]$
λ^*	$0,3448$	$0,2773$	$[-0,1986;\ 0,8883]$

Tabela 4.14: Resultado do ajuste do modelo GTDL.

α	$0,0160$	$0,0137$	$[-0,0108;\ 0,0429]$
β_3	$1,0870$	$0,3270$	$[0,4461;\ 1,7278]$
β_4	$-0,1722$	$0,3853$	$[-0,9274;\ 0,5830]$
β_5	$0,3286$	$0,3007$	$[-0,2608;\ 0,9181]$
β_6	$0,7810$	$0,3203$	$[0,1531;\ 1,4088]$
β_7	$-3,1425$	$0,3728$	$[-3,8731;\ -2,4118]$
β_8	$-2,9835$	$0,3398$	$[-3,6496;\ -2,3175]$
β_9	$-2,5173$	$0,3335$	$[-3,1701;\ -1,8636]$
β_{10}	$-1,7899$	$0,3565$	$[-2,4886;\ -1,0912]$
β_{11}	$-0,2328$	$0,1305$	$[-0,4885;\ 0,0229]$
β_{12}	$0,7967$	$0,2477$	$[0,3112;\ 1,2821]$
β_{13}	$0,1293$	$0,1693$	$[-0,2025;\ 0,4611]$
β_{14}	$-0,3660$	$0,1096$	$[-0,5809;\ -0,1512]$
β_{15}	$-0,5240$	$0,1149$	$[-0,7493;\ -0,2987]$
β_{16}	$-0,2889$	$0,1639$	$[-0,6102;\ 0,0324]$
β_{17}	$0,5201$	$0,1360$	$[0,2534;\ 0,7867]$
β_{18}	$-0,2357$	$0,1871$	$[-0,6024;\ 0,1310]$
β_{19}	$0,1320$	$0,1097$	$[-0,0829;\ 0,3470]$
λ^*	$0,0782$	$0,2160$	$[-0,3452;\ 0,5016]$

Notamos que os β_{18} e β_{19} (que representa a covariável fumante) não são significativos, em ambos os modelos, isto confirma o resultado do método de seleção de covariáveis que indicou que esta covariável é não significativa. O parâmetro θ^* é significativo o que indica que existe heterogeneidade não observada e que covariáveis importantes para explicar o tempo até a morte dos pacientes não foram coletados. O parâmetro α é significativo e positivo, no modelo GTDL com fragilidade, então temos que o tempo tem o efeito de aumentar o risco do paciente assim antecipando a ocorrência da falha, o mesmo não é significativo no modelo GTDL.

Como observamos pela Tabela 4.13 que o parâmetro θ^* que mede a heterogeneidade dos indivíduos é significativo, com isso a utilização do modelo GTDL pode levar

a vícios na estimação do parâmetros. Comparando os valores apresentados pela Tabela 4.14 vemos que as diferenças entre as estimativas pontuais dos modelo GTDL e GTDL com fragilidade é razoável, isto pode ser explicado pelo vício na estimação dos parâmetros quando usamos o modelo GDTL.

Existem vários métodos de seleção de modelos utilizando inferência clássica, os critérios de seleção AIC e BIC são dois deles, e são dada por

$$\begin{aligned} AIC &= -2log(L) + 2k \\ BIC &= -2log(L) + klog(n) \end{aligned}$$

sendo L o máximo da função de verossimilhança, k a quantidade de parâmetros do modelo e n a quantidade de observações. Para mais detalhes sobre os critérios AIC e BIC ver Akaike (1974) e Schwarz (1978), respectivamente. O melhor modelo é o que apresentar o menor valor de AIC e BIC.

Utilizando os critérios AIC e BIC, para comparar os modelos GTDL e GTDL com fragilidade, obtemos os resultados que são apresentados na Tabela 4.15.

Tabela 4.15: Resultado dos critérios de seleção de modelos.

	log verossimilhança	$-2log$ verossimilhança	AIC	BIC
GTDL com fragilidade	$-1596,52$	$3193,04$	$3233,04$	$3325,49$
GTDL	$-1613,19$	$3226,38$	$3264,38$	$3352,24$

Vemos que os valores AIC e BIC são menores para o modelo GTDL com fragilidade, portanto ambos os critérios AIC e BIC indicam que este é o melhor modelo.

4.3 Considerações Finais

Neste Capítulo apresentamos dois métodos de geração de dados do modelo GTDL com fragilidade, um quando não se têm censura e o outro quando existe dados censurados. A partir destes métodos de geração conseguimos calcular a probabilidade de cobertura, o vício e o erro quadrático médio em três situações (sem censura e com 10% e 30%

de censura) e este estudo mostrou que a probabilidade de cobertura para tamanho de amostra razoável fica proxima da nominal e que as estimativas são não viciadas em todas as situações estudadas. Da aplicação com o conjunto de dados reais obtivemos que o modelo com fragilidade se ajustou melhor aos dados do que o modelo sem fragilidade.

Capítulo 5

Análise Bayesiana

Neste Capítulo apresentamos uma aplicação com dados gerados e uma outra com dados reais. Analisamos um conjunto de dados reais, ajustando os modelos GTDL e GTDL com fragilidade e selecionando o melhor modelo. Com um conjunto de dados de tempos multivariados artificial ajustamos o modelo GTDL com fragilidade sob duas abordagens.

Para o mesmo conjunto de dados simulados utilizados na Seção 3.3, apresentamos na Seção 5.1 uma abordagem bayesiana. O ajuste dos modelos GTDL e GTDL com fragilidade em um conjunto de dados reais juntamente com o resultado de um critério de seleção de modelos estão apresentados na Seção 5.2. O modelo GTDL com fragilidade sob duas abordagens se encontra na Seção 5.3.

5.1 Aplicação Utilizando Dados Gerados

Usamos aqui o mesmo conjunto de dados que foi descrito e utilizado na Seção 3.3, utilizamos o modelo GTDL com fragilidade e especificamos distribuições a *priori* independentes para α, β, λ e θ, em que a distribuição a *priori* conjunta é dada por,

$$\pi(\theta, \alpha, \beta, \lambda) = \pi(\theta)\pi(\alpha)\pi(\beta)\pi(\lambda),$$

sendo que,

$$\alpha \text{ e } \beta \sim N(\mu = 0, \sigma^2 = 10^2) \text{ e}$$
$$\lambda \text{ e } \theta \sim Gama(\alpha = 1, \beta = 0,1),$$

utilizamos a parametrização da distribuição gama da forma que foi definida em Mood *et al.* (1974), implicando $E(\lambda) = E(\theta) = 10$ e $Var(\lambda) = Var(\theta) = 100$.

Para determinar os valores dos hiperparâmetros das distribuições a *priori* foi realizado análise de sensibilidade, constatamos que o aumento da variabilidade das distribuição a *priori* não afetou os resumos das distribuições a *posteriori*.

Considerando a função de verossimilhança dada em (3.12) e a distribuição a *priori*, a distribuição a *posteriori* conjunta é dada por

$$\begin{aligned}\pi(\alpha,\beta,\lambda,\theta|\text{dados}) &\propto \left\{\prod_{i=1}^{n}\left(\frac{\lambda\exp(\alpha t_i + x_i'\beta)}{1+\exp(\alpha t_i + x_i'\beta)}\right)^{\delta_i}\left[1+\frac{\lambda\theta}{\alpha}\log\left(\frac{1+\exp(\alpha t_i + x_i'\beta)}{1+\exp(x_i'\beta)}\right)\right]^{-1/\theta-\delta_i}\right\} \\ &\times \exp\left(-\frac{1}{200}\alpha^2\right)\exp\left(-\frac{1}{200}\beta^2\right)\exp(-0,1\lambda)\exp(-0,1\theta). \qquad (5.1)\end{aligned}$$

As distribuições a *posteriori* condicionais são,

$$\begin{aligned}\pi(\alpha|\beta,\lambda,\theta,\text{dados}) &\propto \left\{\prod_{i=1}^{n}\left(\frac{\lambda\exp(\alpha t_i + x_i'\beta)}{1+\exp(\alpha t_i + x_i'\beta)}\right)^{\delta_i}\left[1+\frac{\lambda\theta}{\alpha}\log\left(\frac{1+\exp(\alpha t_i + x_i'\beta)}{1+\exp(x_i'\beta)}\right)\right]^{-1/\theta-\delta_i}\right\} \\ &\times \exp\left(-\frac{1}{200}\alpha^2\right); \qquad (5.2)\end{aligned}$$

$$\begin{aligned}\pi(\beta|\alpha,\lambda,\theta,\text{dados}) &\propto \left\{\prod_{i=1}^{n}\left(\frac{\lambda\exp(\alpha t_i + x_i'\beta)}{1+\exp(\alpha t_i + x_i'\beta)}\right)^{\delta_i}\left[1+\frac{\lambda\theta}{\alpha}\log\left(\frac{1+\exp(\alpha t_i + x_i'\beta)}{1+\exp(x_i'\beta)}\right)\right]^{-1/\theta-\delta_i}\right\} \\ &\times \exp\left(-\frac{1}{200}\beta^2\right); \qquad (5.3)\end{aligned}$$

$$\pi(\lambda|\alpha,\beta,\theta,\text{dados}) \propto \left\{\prod_{i=1}^{n}\lambda^{\delta_i}\left[1+\frac{\lambda\theta}{\alpha}\log\left(\frac{1+\exp(\alpha t_i + x_i'\beta)}{1+\exp(x_i'\beta)}\right)\right]^{-1/\theta-\delta_i}\right\}\exp\left(-0,1\lambda\right); \qquad (5.4)$$

$$\pi(\theta|\alpha,\beta,\lambda,\text{dados}) \propto \left\{\prod_{i=1}^{n}\left[1+\frac{\lambda\theta}{\alpha}\log\left(\frac{1+\exp(\alpha t_i + x_i'\beta)}{1+\exp(x_i'\beta)}\right)\right]^{-1/\theta-\delta_i}\right\}\exp\left(-0,1\theta\right). \qquad (5.5)$$

Fazendo o uso do pacote estatístico R Development Core Team (2008) geramos uma amostra de tamanho 105.000 utilizando o método MCMC, ou seja, consideramos o

algoritmo de Metropolis-Hastings, apresentado no Apêndice B. Adotamos um *burn-in* de 5.000, contornamos o problema de correlação usando saltos de 10. A amostra obtida é de tamanho 10.000 e com o pacote CODA (Plummer *et al.*, 2009) verificamos a convergência utilizando o método de Geweke, que é apresentado no Apêndice C.

A representação gráfica para a verificação da convergência se encontra na Figura 5.1.

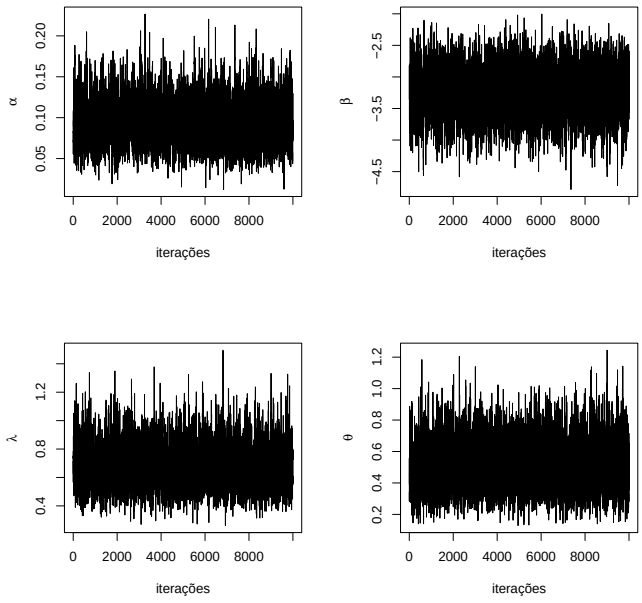

Figura 5.1: Traços das cadeias geradas.

Utilizando o critério de Geweke obtemos os valores que estão na Tabela 5.1. Notamos que os valores estão dentro da banda, considerada de convergência.

Tabela 5.1: Resultados do critério de Geweke.

Parâmetros	Geweke
α	$-0,8205$
β	$0,2789$
λ	$-0,0769$
θ	$-1,0488$

A distribuição a *posteriori* dos parâmetros é apresentada na Figura 5.2, onde notamos que existe assimetria nas densidades.

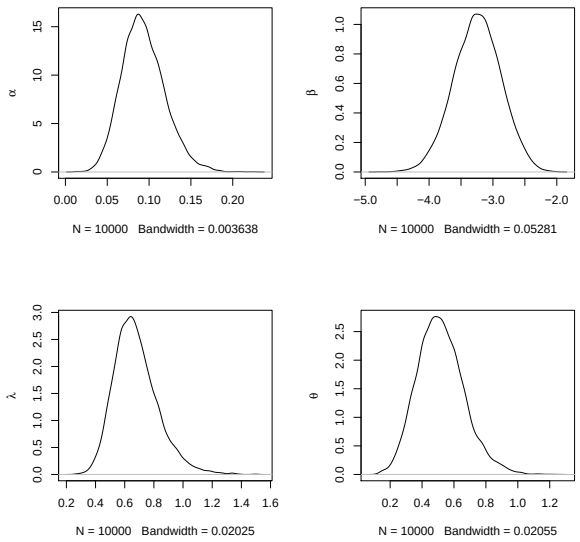

Figura 5.2: Distribuições a *posteriori*

A Tabela 5.2 mostra as estatísticas resumos a *posteriori* para os parâmetros, tais como média, mediana, desvio padrão (DP) e intervalo de credibilidade (IC), que foi construído considerando os percentis 2,5% e 97,5%. Note que a média e a mediana estão próximos aos valores verdadeiros dos parâmetros.

Tabela 5.2: Resumos a *posteriori*.

Parâmetros	Média	Mediana	DP	IC (95%)
α	$0,1031$	$0,1013$	$0,0297$	$[0,0509;\ 0,1657]$
β	$-3,0890$	$-3,0870$	$0,4467$	$[-3,9680;\ -2,2330]$
λ	$0,2963$	$0,2881$	$0,0731$	$[0,1778;\ 0,4609]$
θ	$0,6186$	$0,6166$	$0,1814$	$[0,3651;\ 1,0820]$

Observando a Tabela 5.2 vemos que a mediana das distribuições se aproxima mais das estimativas pontuais do que a média, isso pode ser justificado devido a assimetria que existe nas distribuições a *posteriori*. Percebemos ainda que os intervalo de credibilidade construído cobrem os verdadeiros valores dos parâmetros.

Fazendo uma comparação entre as Tabela 3.1 e Tabela 5.2 notamos que as estimativas pontuais estão próximas, os desvios padrões na inferência bayesiana são menores, conseqüentemente os intervalos de credibilidade são menores do que os intervalos de confiança.

5.2 Aplicação Utilizando Dados Reais

Utilizamos aqui o mesmo conjunto de dados que foi descrito e utilizado na Seção 4.2.

Para o modelo GTDL com fragilidade especificamos distribuições a *priori* independentes para α, β_3, ..., β_{19}, λ e θ, em que a distribuição a *priori* conjunta é dada por,

$$\pi(\theta, \alpha, \boldsymbol{\beta}, \lambda) = \pi(\theta)\pi(\alpha)\pi(\beta_3)\ldots\pi(\beta_{19})\pi(\lambda).$$

sendo que,

$$\alpha,\, \beta_3, \ldots, \beta_{19} \sim N(\mu = 0, \sigma^2 = 10^2)$$
$$\lambda \text{ e } \theta \sim Gama(\alpha = 1, \beta = 0,1).$$

Utilizamos a parametrização da distribuição gama da forma que foi definida em Mood *et al.* (1974), implicando $E(\lambda) = E(\theta) = 10$ e $Var(\lambda) = Var(\theta) = 100$. Com isso teremos as mesmas *posteriori* conjunta e marginal apresentadas pelas equações (5.1), (5.2), (5.3), (5.4) e (5.5), apenas com a diferença na distribuição condicional para β_i que é dada por

$$\begin{aligned}\pi(\beta_i|\beta_{-i}, \alpha, \lambda, \theta, \text{dados}) &\propto \left\{\prod_{i=1}^{n}\left(\frac{\lambda\exp(\alpha t_i + \mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\alpha t_i + \mathbf{x}_i'\boldsymbol{\beta})}\right)^{\delta_i}\left[1 + \frac{\lambda\theta}{\alpha}\log\left(\frac{1+\exp(\alpha t_i + \mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right)\right]^{-1/\theta-\delta_i}\right\} \\ &\times \exp\left(-\frac{1}{200}\beta_i^2\right);\end{aligned}$$

sendo que $\boldsymbol{\beta_{-i}}$ significa $(\beta_3, \ldots, \beta_{i-1}, \beta_{i+1}, \beta_{19})$, para $i = 3, \ldots, 19$.

A mesma forma de especificação de distribuições a *priori* apresentada para o modelo GTDL com fragilidade foi feito para o modelo GTDL apenas com a ausência da distribuição a *priori* para o parâmetro θ.

Considerando a função de verossimilhança dada em (2.4) e as distribuições a *priori*, a distribuição a *posteriori* conjunta é dada por

$$\begin{aligned}\pi(\alpha, \beta, \lambda|\text{dados}) &\propto \left\{\prod_{i=1}^{n}\left(\lambda\frac{\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}{1+\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}\right)^{\delta_i}\left(\frac{1+\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)^{-\lambda/\alpha}\right\} \\ &\times \exp\left(-\frac{\alpha^2}{200}\right)\exp\left(-\frac{\beta_3^2}{200}\right)\times\ldots\times\exp\left(-\frac{\beta_{19}^2}{200}\right)\times\exp(-0,1\lambda).\end{aligned}$$

As distribuições a *posteriori* condicionais são,

$$\pi(\alpha|\boldsymbol{\beta}, \lambda, \text{dados}) \propto \left\{\prod_{i=1}^{n}\left(\lambda\frac{\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}{1+\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}\right)^{\delta_i}\left(\frac{1+\exp(\alpha t + \mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)^{-\lambda/\alpha}\right\}\exp\left(-\frac{\alpha^2}{200}\right);$$

$$\pi(\beta_i|\alpha,\boldsymbol{\beta_{-i}},\lambda,\text{dados}) \propto \left\{\prod_{i=1}^{n}\left(\lambda\frac{\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}\right)^{\delta_i}\left(\frac{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)^{-\lambda/\alpha}\right\}\exp\left(-\frac{\beta_i^2}{200}\right);$$

$$\pi(\lambda|\alpha,\boldsymbol{\beta},\text{dados}) \propto \left\{\prod_{i=1}^{n}\left(\lambda\frac{\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}\right)^{\delta_i}\left(\frac{1+\exp(\alpha t+\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})}\right)^{-\lambda/\alpha}\right\}\exp(-0,1\lambda).$$

Utilizando o pacote estatístico R Development Core Team (2008) geramos uma amostra de tamanho 105.000 utilizando o algoritmo de Metropolis-Hastings. Adotamos um *burn-in* de 5.000, contornamos o problema de correlação usando saltos de 10. A amostra obtida é de tamanho 10.000 e com o pacote CODA (Plummer *et al.*, 2009) verificamos a convergência, isto para ambos os modelos.

As estatísticas do teste de Geweke são apresentadas nas Tabelas 5.4 e 5.3, para os modelos GTDL e GTDL com fragilidade, respectivamente, notamos pelas mesma que as estatísticas estão dentro da banda considerada de convergência.

Tabela 5.3: Resultados do critério de Geweke para o modelo GTDL com fragilidade.

Parâmetros	Geweke	Parâmetros	Geweke
β_3	$1,0642$	β_{13}	$0,5238$
β_4	$0,9569$	β_{14}	$0,3627$
β_5	$0,0754$	β_{15}	$0,2964$
β_6	$0,3184$	β_{16}	$0,9986$
β_7	$0,9963$	β_{17}	$0,5175$
β_8	$0,9943$	β_{18}	$-1,1898$
β_9	$0,3608$	β_{19}	$0,3545$
β_{10}	$-0,0787$	α	$0,1741$
β_{11}	$-0,2394$	λ	$0,3637$
β_{12}	$-0,0035$	θ	$0,2275$

Tabela 5.4: Resultados do critério de Geweke para o modelo GTDL.

Parâmetros	Geweke	Parâmetros	Geweke
β_3	$-0,2411$	β_{13}	$-1,5159$
β_4	$-1,0060$	β_{14}	$-0,5359$
β_5	$-0,1894$	β_{15}	$0,8397$
β_6	$-0,0912$	β_{16}	$-1,5727$
β_7	$-1,7506$	β_{17}	$-0,4698$
β_8	$-1,6010$	β_{18}	$-1,7322$
β_9	$-1,2867$	β_{19}	$-1,0568$
β_{10}	$-1,9465$	α	$-0,0128$
β_{11}	$-1,4311$	λ	$-0,1866$
β_{12}	$-0,6451$		

Apresentamos nas Tabelas 5.6 e 5.5 alguns resumos a *posteriori* como, média, mediana, desvio padrão (DP) e intervalo de credibilidade (IC) que foi construído com 95%, isso para os modelos GTDL e GTDL com fragilidade, respectivamente.

Tabela 5.5: Resultados do ajuste do modelo GTDL com fragilidade.

Parâmetros	Média	Mediana	DP	IC(95%)
β_3	$1,7480$	$1,7380$	$0,4937$	$[0,7817;\ 2,7520]$
β_4	$-0,1627$	$-0,1675$	$0,5327$	$[-1,2516;\ 0,9351]$
β_5	$0,7664$	$0,7624$	$0,4652$	$[-0,1460;\ 1,7151]$
β_6	$1,2010$	$1,1970$	$0,4873$	$[0,2282;\ 2,1936]$
β_7	$-4,1080$	$-4,1070$	$0,6039$	$[-5,3390;\ -2,8882]$
β_8	$-3,9590$	$-3,9710$	$0,5969$	$[-5,1318;\ -2,7572]$
β_9	$-3,1640$	$-3,1670$	$0,5717$	$[-4,2998;\ -2,0158]$
β_{10}	$-2,0200$	$-2,0360$	$0,6172$	$[-3,2031;\ -0,7553]$
β_{11}	$-0,1766$	$-0,1768$	$0,2133$	$[-0,6116;\ 0,2422]$
β_{12}	$1,1100$	$1,1080$	$0,3856$	$[0,3552;\ 1,8879]$
β_{13}	$0,0235$	$0,0256$	$0,2606$	$[-0,5046;\ 0,5551]$
β_{14}	$-0,4613$	$-0,4617$	$0,1764$	$[-0,8153;\ -0,1026]$
β_{15}	$-0,7670$	$-0,7672$	$0,1947$	$[-1,1631;\ -0,3727]$
β_{16}	$-0,3104$	$-0,3090$	$0,2648$	$[-0,8474;\ 0,2209]$
β_{17}	$0,6418$	$0,6393$	$0,2207$	$[0,1996;\ 1,0924]$
β_{18}	$-0,1307$	$-0,1298$	$0,2987$	$[-0,7349;\ 0,4774]$
β_{19}	$0,0812$	$0,0796$	$0,1823$	$[-0,2806;\ 0,4370]$
α	$0,1400$	$0,1398$	$0,0333$	$[0,0728;\ 0,2068]$
θ	$0,6209$	$0,6051$	$0,1491$	$[0,3685;\ 0,9645]$
λ	$1,4580$	$1,4420$	$0,4323$	$[0,6515;\ 2,3720]$

Tabela 5.6: Resultados do ajuste do modelo GTDL.

Parâmetros	Média	Mediana	DP	IC(95%)
β_3	$1,0980$	$1,1010$	$0,3598$	$[0,3783;\ 1,8146]$
β_4	$-0,1898$	$-0,1834$	$0,4040$	$[-1,0064;\ 0,6170]$
β_5	$0,3293$	$0,3283$	$0,3301$	$[-0,3423;\ 0,9940]$
β_6	$0,7850$	$0,7863$	$0,3531$	$[0,08156;\ 1,4859]$
β_7	$-3,1560$	$-3,1520$	$0,3918$	$[-3,9558;\ -2,3941]$
β_8	$-2,9940$	$-2,9940$	$0,3713$	$[-3,7494;\ -2,2557]$
β_9	$-2,5170$	$-2,5120$	$0,3663$	$[-3,2542;\ -1,7886]$
β_{10}	$-1,7650$	$-1,7670$	$0,3880$	$[-2,5231;\ -0,9700]$
β_{11}	$-0,2320$	$-0,2323$	$0,1502$	$[-0,5289;\ 0,0787]$
β_{12}	$0,8036$	$0,8029$	$0,2739$	$[0,2455;\ 1,3541]$
β_{13}	$0,1308$	$0,1298$	$0,1895$	$[-0,2460;\ 0,5238]$
β_{14}	$-0,3675$	$-0,0368$	$0,1272$	$[-0,6187;\ -0,1066]$
β_{15}	$-0,5268$	$-0,5264$	$0,1293$	$[-0,7933;\ -0,2690]$
β_{16}	$-0,2880$	$-0,2880$	$0,1813$	$[-0,6604;\ 0,0743]$
β_{17}	$0,5230$	$0,5222$	$0,1535$	$[0,2174;\ 0,8322]$
β_{18}	$-0,2402$	$-0,2386$	$0,2055$	$[-0,6603;\ 0,1758]$
β_{19}	$0,1331$	$0,1340$	$0,1252$	$[-0,1239;\ 0,3838]$
α	$0,0160$	$0,0161$	$0,0160$	$[-0,0167;\ 0,0480]$
λ	$1,0810$	$1,0870$	$0,2602$	$[0,5753;\ 1,6375]$

Utilizando as Tabelas 5.6, 5.5 e **??** notamos que as estimativas clássicas e bayesiana são próximas, este fato já era esperado pois utilizamos distribuições a *priori* de tal maneira que não acrescenta-se informação.

Existem vários critérios de seleção de modelos quando utilizamos inferência bayesiana, um deles é o *deviance information criterion* (DIC), que é calculado da forma

$$DIC = 2\widehat{D}_{avg}(y) - D_{\widehat{\psi}}(y),$$

sendo que ψ representa os parâmetros do modelo e

$$D(y, \psi) = -2logp(y|\psi),$$

quando $p(y|\psi)$ representa a função de densidade de probabilidade e

$$D_{\widehat{\psi}}(y) = D(y, \widehat{\psi}),$$

cujo $\widehat{\psi}$ é a média a *posteriori* dos parâmetros e

$$\widehat{D}_{avg} = \frac{1}{L}\sum_{l=1}^{L} D(y, \psi^l),$$

sendo que L representa a quantidade de observações da cadeia e l representa a iteração da cadeia. O modelo que se adequou melhor aos dados é o que apresentar menor DIC. Para mais detalhes do critério DIC ver Gelman *et al.* (2004).

Para o conjunto de dados reais, calculamos o DIC para os modelos GTDL e GTDL com fragilidade, o resultado é mostrado na Tabela 5.7.

Tabela 5.7: Resultado do critério DIC.

Modelo	DIC
GTDL com fragilidade	2275,448
GTDL	2782,114

Observamos na Tabela 5.7 que o modelo com menor DIC é o modelo GTDL com fragilidade conseqüentemente o melhor modelo entre os dois, este resultado coincide com os critérios AIC e BIC quando utilizando inferência clássica.

5.3 O Modelo GTDL com Fragilidade Sob Duas Abordagens

Nesta Seção comparamos dois métodos de estimação do modelo GTDL com fragilidade usando um conjunto de dados com eventos recorrentes.

O primeiro modelo é apresentado pelas funções de risco, de sobrevivência e de verossimilhança dadas em (3.9), (3.8) e (3.12), respectivamente. O segundo modelo é apresentado pelas funções de risco (3.3), a correspondente função de sobrevivência (3.4) e a função de verossimilhança é construída a partir da equação dada em (1.8). Chamamos o primeiro modelo por modelo marginalizado e o segundo por modelo completo. Para evitar a falta de identificabilidade adotamos $\lambda = 1$ no modelo completo.

A diferença entre os dois modelos é que o modelo marginalizado não estima os valores da fragilidade individual mas estima a variância das fragilidades, conseqüentemente este modelo possui menos parâmetros a serem estimados do que o modelo completo, que por sua vez estima as fragilidades individuais.

Para a comparação geramos uma amostra com eventos recorrentes, ou seja, para cada indivíduo é observado mais que um tempo de ocorrência do evento de interesse. A geração dos tempos é feita da seguinte forma:

1. adotamos que a quantidade de indivíduos é igual a 15 e que para cada indivíduo são observados 3 tempos;
2. fixamos $\beta = -3,00$, $\alpha = 0,10$, $\lambda = 1,00$ e $\theta = 0,50$;
3. geramos v_i da distribuição gama($1/\theta$,$1/\theta$), $i = 1, ..., 15$;
4. gerarmos u_{i1}, u_{i2} e u_{i3} da distribuição $U(0, 1)$;
5. geramos x_{i1}, x_{i2} e x_{i3} da distribuição Bernoulli com probabilidade de sucesso igual a 0,50;
6. geramos o tempo t_{iz} a partir da substituição das quantidades u_{iz}, v_i, β, α, λ e x_{iz} na equação (4.2), para $z = 1, 2, 3$;
7. Fixamos $\delta_{iz} = 1$, pois os tempos são não censurados.

A distribuição a *priori* para o modelo completo é a mesma da utilizada na Seção 5.3.1, conseqüentemente a distribuição a *posteriori* conjunta é igual a dada em (5.1). Para o modelo completo especificamos distribuições a *priori* independentes para α, β e v_i, em que a distribuição a *priori* conjunta é dada por,

$$\pi(\alpha, \beta, \boldsymbol{v}) = \pi(\alpha)\pi(\beta)\pi(v_1)\ldots\pi(v_{15}),$$

sendo que,

$$\alpha \text{ e } \beta \sim N(\mu = 0, \sigma^2 = 10^2)$$
$$v_i \sim Gama(\alpha = 2, \beta = 2).$$

Utilizando as funções de risco e de sobrevivência dadas em (3.3) e (3.4), respectivamente, a fórmula da função de verossimilhança dada em (1.8) e a distribuição a *priori* conjunta, a distribuição a *posteriori* conjunta para o modelo completo é dado por

$$\begin{aligned}\pi(\alpha, \beta, \boldsymbol{v}|dados) &= \prod_{i=1}^{15}\left[\prod_{j=1}^{3}\left(\frac{v_i \exp(\alpha t_{ij} + x_{ij}\beta)}{1 + \exp(x_{ij}\beta)}\right)^{\delta_{ij}}\left(\frac{1 + \exp(\alpha t_{ij} + x_{ij}\beta)}{1 + \exp(x_{ij}\beta)}\right)^{-v_i/\alpha}\right] \\ &\times \exp\left(-\frac{\alpha}{200}\right) \times \exp\left(-\frac{\beta}{200}\right) \times v_1^{0,5}\exp(-0,5v_1) \times ... \times v_{15}^{0,5}\exp(-0,5v_{15}).\end{aligned} \tag{5.6}$$

As distribuições a *posteriori* são

$$\begin{aligned}\pi(\alpha|dados, \beta, \boldsymbol{v}) &= \prod_{i=1}^{15}\left[\prod_{j=1}^{3}\left(\frac{v_i \exp(\alpha t_{ij} + x_{ij}\beta)}{1 + \exp(x_{ij}\beta)}\right)^{\delta_{ij}}\left(\frac{1 + \exp(\alpha t_{ij} + x_{ij}\beta)}{1 + \exp(x_{ij}\beta)}\right)^{-v_i/\alpha}\right] \\ &\times \exp\left(-\frac{\alpha}{200}\right);\end{aligned} \tag{5.7}$$

$$\pi(\beta|dados,\alpha,\boldsymbol{v}) = \prod_{i=1}^{15}\left[\prod_{j=1}^{3}\left(\frac{v_i\exp(\alpha t_{ij}+x_{ij}\beta)}{1+\exp(x_{ij}\beta)}\right)^{\delta_{ij}}\left(\frac{1+\exp(\alpha t_{ij}+x_{ij}\beta)}{1+\exp(x_{ij}\beta)}\right)^{-v_i/\alpha}\right] \times \exp\left(-\frac{\beta}{200}\right); \quad (5.8)$$

$$\pi(v_k|dados,\alpha,\beta,v_{-k}) = \prod_{i=1}^{15}\left[\prod_{j=1}^{3}\left(\frac{v_i\exp(\alpha t_{ij}+x_{ij}\beta)}{1+\exp(x_{ij}\beta)}\right)^{\delta_{ij}}\left(\frac{1+\exp(\alpha t_{ij}+x_{ij}\beta)}{1+\exp(x_{ij}\beta)}\right)^{-v_i/\alpha}\right] \times v_k^{0,5}\exp(-0,5v_k). \quad (5.9)$$

De maneira análoga a adotada na Seção 5.3.1 obtemos uma amostra de tamanho 10.000 e a partir desta as estatísticas: média, desvio padrão (DP), intervalo de credibilidade de 95% (IC (95%)) e teste de Geweke estão apresentadas nas Tabelas 5.8 e 5.9 para os modelos marginalizado e completo, respectivamente.

Tabela 5.8: Resumos a *posteriori* para o modelo marginalizado.

Parâmetros	Média	DP	IC (95%)	Geweke	Valor Verdadeiro
α	$0,1002$	$0,0497$	$[0,0137;\ 0,2092]$	-0,7149	0,10
β	$-3,1050$	$0,4283$	$[-3,9405;\ -2,2699]$	1,6027	-3,00
λ	$1,2743$	$0,2769$	$[0,8270;\ 1,9100]$	0,3689	1,00
θ	$0,3146$	$0,2081$	$[0,0412;\ 0,8204]$	-0,8668	0,50

Tabela 5.9: Resumos a *posteriori* para o modelo completo.

Parâmetros	Média	DP	IC (95%)	Geweke	Valor Verdadeiro
α	$0,2895$	$0,0661$	$[0,1656;\ 0,4240]$	-0,6560	0,1000
β	$-4,2990$	$0,5127$	$[-5,3397;\ -3,3278]$	1,4964	-3,0000
v_1	$1,8030$	$0,7127$	$[0,6326;\ 3,3855]$	-1,4122	2,2371
v_2	$3,6380$	$1,6013$	$[1,0170;\ 7,2606]$	0,9464	1,4617
v_3	$1,8920$	$0,7156$	$[0,6665;\ 3,4264]$	-0,9324	2,0670
v_4	$1,0120$	$0,4048$	$[0,3220;\ 1,8742]$	1,6184	0,4673
v_5	$1,0060$	$0,3773$	$[0,3533;\ 1,8104]$	1,0443	0,9414
v_6	$2,9230$	$1,1861$	$[0,9336;\ 5,4623]$	0,6318	0,7377
v_7	$1,5560$	$0,5886$	$[0,5469;\ 2,8042]$	0,0944	0,2937
v_8	$1,3610$	$0,5190$	$[0,4601;\ 2,4647]$	-0,0774	0,7768
v_9	$0,9071$	$0,3391$	$[0,3078;\ 1,6240]$	-1,3603	0,3472
v_{10}	$0,1417$	$0,0557$	$[0,0456;\ 0,2591]$	-0,0644	1,7833
v_{11}	$0,2434$	$0,0919$	$[0,0824;\ 0,4408]$	-0,3366	0,6953
v_{12}	$0,2206$	$0,0879$	$[0,0706;\ 0,4109]$	0,6590	2,5111
v_{13}	$1,5750$	$0,6319$	$[0,4904;\ 2,9394]$	-0,5200	1,0989
v_{14}	$3,7730$	$1,4274$	$[1,2887;\ 6,8251]$	-0,0759	0,5943
v_{15}	$1,8440$	$0,7159$	$[0,6040;\ 3,3582]$	0,1941	2,2074

Notamos pelas Tabelas 5.8 e 5.9 que as estimativas de α e de β do modelo marginalizado ficam mais próximas dos valores verdadeiros do que as estimativas obtidas pelo modelo completo.

5.4 Considerações Finais

A metodologia bayesiana aplicada neste Capítulo atingiu resultados parecidos com os resultados obtidos quando utilizamos a metodologia clássica, isso tanto para o conjunto de dados reais quanto gerados, o que já era esperado pois utilizamos distribuições a *priori* com variância grande deixando assim a distribuição a *priori* com pouca informação. A comparação entre os modelos marginalizado e completo mostra que a estimação dos parâmetros α e β do modelo marginalizado foi a que ficou mais próxima do verdadeiro valor, mas não podemos esquecer das diferenças entre os modelos, pois só o modelo completo estima as fragilidades individuais e o modelo marginalizado é o único que estima a variância das fragilidades.

Capítulo 6

Conclusões e Perspectivas

Neste trabalho apresentamos e estudamos o modelo logístico generalizado dependente do tempo com fragilidade, utilizando inferência clássica e bayesiana. Também, mostramos a sua aplicação em um conjunto de dados reais.

Do ponto de vista clássico, não tivemos problema com identificabilidade. Este fato viabilizou o estudo com dados gerados e com isso desenvolvemos o cálculo da probabilidade de cobertura, o vício e o erro quadrático médio, para diferentes níveis de censura.

Sob o enfoque bayesiano, não existiu dificuldades de implementação do método Metropolis-Hastings. Usando o modelo GTDL com fragilidade e para tempos multivariados a grande vantagem esta na utilização da inferência bayesiana pois conseguimos atingir nosso objetivos, ou seja, os resultados alcançados foram satisfatório.

Com os resultados dos estudos de simulação e com a aplicação em dados reais concluímos que o modelo GTDL com fragilidade capta a heterogeneidade não observada das unidades e o valor das covariáveis que ou não foram incorporadas no problema ou que por algum motivo não puderam ser coletadas influenciam na determinação do tempo de falha.

Como perspectiva de continuidade deste trabalho propomos o estudo do custo de adicionar o parâmetro que mede a heterogeneidade no modelo GTDL. Sob o ponto de vista bayesiano também propomos trabalhar com distribuições a *priori* não informativa, como por exemplo, a distribuição a *priori* de referência, proposta por Bernardo (1979).

Referências Bibliográficas

Aalen, O. Heterogeneity in survival analysis. *Statistics in Medicine*, v. 7, p. 1121–1137, 1988.

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

Andersen, P. K.; Gill, R. D. Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, v. 10, p. 1100–1120, 1982.

Aranda-Ordaz, F. J. An extension of the proportional hazards model for grouped data. *Biometrics*, v. 39, p. 109–117, 1983.

Bernardo, J. M. Reference posterior distribution for bayesian inference. *Journal of the Royal Statistics Society B*, v. 41, p. 113–47, 1979.

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*, v. 65, p. 141–151, 1978.

Clayton, D. G. A monte carlo method for bayesian inference in frailty models. *Biometrics*, v. 47, p. 467–485, 1991.

Clayton, D. G.; Cuzick, J. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society. Series A*, v. 148, p. 82–117, 1985.

Colosimo, E. A.; Giolo, S. R. *Análise de Sobrevivência Aplicada*. Edgard Blücher, São Paulo, SP, 2006.

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

Cremasco, C. P. *Modelagem de Dados de Sobrevivência Via Modelo de Risco Logístico Generalizado*. Março 2005, 68. Dissertação (Mestrado), Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, Março 2005.

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

Etezadi-Amoli, J.; Ciampi, A. Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function. *Biometrics*, v. 43, p. 181–192, 1987.

Gamerman, D. *Simulação Estocástica via Cadeias de Markov*. ABE, São Paulo, 1996.

Gamerman, D.; Lopes, H. F. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, Londres, 2nd ed., 2006.

Gelman, A.; Carlin, J. B.; Stern, H. S.; Rubin, D. B. *Bayesian Data Analysis*. CHAPMAN & HALL/CRC, 2nd ed., 2004.

Geweke, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). *In Bayesian Statistics*, v. 4, p. 169–193, 1992.

Gompertz, B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*, v. 115, p. 513–583, 1825.

Ha, I. D.; MacKenzie, G. Robust frailty modelling using non-proportional hazards models. *Statistical Modelling*, v. 10(3), p. 315–332, 2010.

Ha, I. D.; Lee, Y.; Song, J. K. Hierarchical likelihood approach for frailty models. *Biometrika*, v. 88, p. 233–243, 2001.

Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, v. 57, p. 97–109, 1970.

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

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

Hougaard, P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, v. 73, p. 387–396, 1986a.

Hougaard, P. A class of multivariate failure time distributions. *Biometrika*, v. 73, p. 671–678, 1986b.

Ibrahim, J.; Chen, M.; Sinha, D. *Bayesian Survival Analysis*. Springer, New York, 2001.

INCA. Instituto nacional de câncer, 2010. URL `http://www.inca.gov.br/conteudo_view.asp?id=340`. acesso em 20 de Novembro de 2010.

Kalbfleisch, J. F.; Prentice, R. L. *The Statistical Analisys of Failure Time Data*. John Wiley and Sons, New York, NY, 1980.

Kaplan, E. L.; Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53, p. 457–481, 1985.

Korsgaard, I. R.; Madsen, P.; Jensen, J. Bayesian inference in the semiparametric log normal model using gibbs sampling. *Genetics Selection Evolution*, v. 30, p. 241–256, 1998.

Lawless, J. *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons, New York, 1982.

Loprinzi, C.; Laurie, J.; Wieand, H.; Krook, J.; Novotny, P.; Kugler, J.; Bartel, J.; Law, M.; Bateman, M.; Klatt, N.; et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North central cancer treatment group. *Journal of Clinical Oncology*, v. 12(3), p. 601–7, 1994.

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

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

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

Mackenzie, G. On a non-proportional hazards regression model for repeated medical random counts. *Statistics in Medicine*, v. 16, p. 1831–1843, 1997.

Mackenzie, G. A logistic regression model for survival data. *In: 17th International Workshop in Statistical Modelling*, pages 431–438, 2002.

McKeague, I. W.; Sasieni, P. D. A partly parametric additive risk model. *Biometrika*, v. 81, p. 501–514, 1994.

Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, v. 521, p. 1087–1092, 1953.

Mood, A. M.; Graybill, F. A.; Boes, D. C. *Introduction to the Theory of Statistics*. McGraw-Hill, Tokyo, 3th ed., 1974.

Nelson, W. Theory and applications of hazard plotting for censored failure data. *Technometrics*, v. 14, p. 945–965, 1972.

Neuhaus, J. M.; Hauck, W. W.; Kalbfleisch, J. D. The effects of mixture distribution misspecification when fitting mixed-effects logistic. *Biometrika*, v. 79, p. 755–762, 1992.

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

Parreira, D. R. M. *Um Modelo de Risco Proporcional Dependente do Tempo*. Junho 2007. Dissertação (Mestrado), Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, Junho 2007.

Paulino, C. D.; Turkman, M. A. A.; Murteira, B. *Estatística Bayesiana*. Fundação Caloute Gulbenkian, Av. de Berna, Lisboa, 2003.

Plummer, M.; Best, N.; Cowles, K.; Vines, K. *CODA: Output analysis and diagnostics for MCMC*, 2009. R package version 0.13-4.

Prentice, R. L. Linear rank tests with right censored data. *Biometrika*, v. 65, p. 167–179, 1978.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. version 2.8.1 (2008-12-22), ISBN 3-900051-07-0.

Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, v. 6, p. 461–464, 1978.

Tibshirani, R. J.; Ciampi, A. A family of proportional and additive hazard models for survival data. *Biometrics*, v. 39, p. 141–147, 1983.

Tomazella, V. L. D. *Modelagem de Dados de Eventos Recorrentes via Processo de Poison com Termo de Fragilidade*. Tese (Doutorado), ICMC, Instituto de Ciências Matemáticas e Computação, São Carlos, Julho 2003.

Tomazella, V. L. D.; Louzada-Neto, F.; Andrade, M. G. Bayesian modelling for multivariate lifetime data with a homogeneous poisson process with a frailty term. *Brazilian Journal of Probability and Statistics*, v. 18, p. 19–35, 2004.

Vaupel, J. W.; Manton, K. G.; Stallard, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, v. 16, p. 439–454, 1979.

Wilkinson, P. *Lung cancer in Northern Ireland*. 1995. Dissertação (Mestrado), Queen's University of Belfast, Belfast, 1995.

Apêndice A

Detalhes do Ajuste com Dados Reais

Tabela A.1: Informações coletadas.

tempo	numérico	continua
indicador de censura	morte= 1 censura = 0	binária
idade em anos	numérico	continua
sexo	homem= 1 mulher= 0	binária
tratamento	paliativos= 1 cirurgia= 2 quimioterapia= 3 radioterapia= 4 quimioterapia+radioterapia= 5	categórica
status de performance	atividade normal= 1 trabalho leve= 2 incapacidade de trabalho= 3 $> 50\%$ comportamento= 4 de cama= 5	categórica
tipo de célula	escamosa= 1 célula pequena= 2 adeno = 3 outra= 4	categórica
nível de sódio	>= 136mmol/l= 1 < 136mmol/l= 2	categórica
nível de albumina	$>= 35g/l = 1$ $< 35g/l = 2$	categórica
metástases	não= 1 sim= 2 não sabe= 3	categórica
fumante	não= 1 sim= 1 ex= 3	categórica

Tabela A.2: Categorias.

	categoria	código
idade em anos	numérico	β_1
sexo	homem	$\beta_2 = 1$
	mulher	$\beta_2 = 0$
tratamento	paliativos	$\beta_3 = 1, \beta_4 = 0, \beta_5 = 0, \beta_6 = 0$
	cirurgia	$\beta_3 = 0, \beta_4 = 1, \beta_5 = 0, \beta_6 = 0$
	quimioterapia	$\beta_3 = 0, \beta_4 = 0, \beta_5 = 1, \beta_6 = 0$
	radioterapia	$\beta_3 = 0, \beta_4 = 0, \beta_5 = 0, \beta_6 = 1$
	quimioterapia+radioterapia	$\beta_3 = 0, \beta_4 = 0, \beta_5 = 0, \beta_6 = 0$
Status de performance	atividade normal	$\beta_7 = 1, \beta_8 = 0, \beta_9 = 0, \beta_{10} = 0$
	trabalho leve	$\beta_7 = 0, \beta_8 = 1, \beta_9 = 0, \beta_{10} = 0$
	incapacidade de trabalho	$\beta_7 = 0, \beta_8 = 0, \beta_9 = 1, \beta_{10} = 0$
	>50% comportamento	$\beta_7 = 1, \beta_8 = 0, \beta_9 = 0, \beta_{10} = 1$
	de cama	$\beta_7 = 0, \beta_8 = 0, \beta_9 = 0, \beta_{10} = 0$
tipo de célula	escamosa	$\beta_{11} = 1, \beta_{12} = 0, \beta_{13} = 0$
	célula pequena	$\beta_{11} = 1, \beta_{12} = 1, \beta_{13} = 0$
	adeno	$\beta_{11} = 1, \beta_{12} = 0, \beta_{13} = 1$
	outra	$\beta_{11} = 1, \beta_{12} = 0, \beta_{13} = 0$
nível de sódio	$>= 136 mmol/l$	$\beta_{14} = 1$
	$< 136 mmol/l$	$\beta_{14} = 0$
nível de albumina	$>= 35 g/l$	$\beta_{15} = 1$
	$< 35 g/l$	$\beta_{15} = 0$
Metástases	não	$\beta_{16} = 1, \beta_{17} = 0$
	sim	$\beta_{16} = 0, \beta_{17} = 1$
	não se sabe	$\beta_{16} = 0, \beta_{17} = 0$
fumante	não	$\beta_{18} = 1, \beta_{19} = 0$
	sim	$\beta_{18} = 0, \beta_{19} = 1$
	ex	$\beta_{18} = 0, \beta_{19} = 0$

Apêndice B

Metropolis-Hastings

O algoritmo de Metropolis-Hastings é o resultado da proposta de Metropolis *et al.* (1953) e da melhora sugerida por Hastings (1970). Este método de geração só é indicado quando a expressão da distribuição que queremos gerar não é conhecida e os outros métodos de geração considerados exatos não funcionam.

Suponha que temos uma distribuição de equilíbrio $\pi(\theta)$, esta pode ser a *posteriori*. A idéia central do algoritmo é a construção de uma cadeia, cujos valores são amostras da distribuição $\pi(\theta)$. A geração de um possível valor de θ é feita a partir de uma distribuição diferente da $\pi(\theta)$, mas este valor é aceito para a cadeia com uma certa probabilidade; caso a cadeia convirja, o sistema de correção (onde o valor é aceito ou não) é o que garante que a distribuição de equilíbrio $\pi(\theta)$ seja a distribuição limite.

Com um pouco mais de formalidade reescrevemos a idéia do algoritmo. Suponha que a cadeia esteja na posição θ^j e um valor θ^{j+1} foi gerado da distribuição $\gamma(.|\theta)$. Observe que a geração do θ^{j+1} pode depender do estado da cadeia j. Um exemplo simples deste fato é quando utilizamos como geradora de candidato o chamado passeio aleatório. O valor θ^{j+1} é aceito com probabilidade

$$\alpha(\theta^j, \theta^{j+1}) = min\left(1, \frac{\pi(\theta^{j+1})\gamma(\theta^{j+1}, \theta^j)}{\pi(\theta^j)\gamma(\theta^j, \theta^{j+1})}\right), \tag{B.1}$$

lembrando que $\pi(.)$ é a distribuição de equilíbrio.

Agora que já descrevemos a idéia vamos expor os passos do algoritmo,

1) iniciar o contador de iterações da cadeia $k = 1$ e atribuir um valor para θ^1;

2) gerar um candidato a θ^{k+1} da distribuição $\gamma(\theta)$;

3) calcular a probabilidade de aceitação $\alpha(\theta^k, \theta^{k+1})$ utilizando a equação (B.1);

4) gerar uma valor u da distribuição $U(0, 1)$;

5) se o valor u for menor que $\alpha(\theta^k, \theta^{k+1})$ então o candidato θ^{k+1} entra na cadeia na posição $k + 1$ caso contrário, o valor θ^{k+1} da cadeia é o mesmo que em θ^k e mude o contador k para $k + 1$;

6) repetir os passos *2* até *5*.

Existem outros métodos de simulação via MCMC, como por exemplo, Metropolis, Gibbs Sampling (para mais detalhes ver Gamerman e Lopes, 2006).

Apêndice C

Método de Geweke

A descrição mais completa do método pode ser encontrada em Geweke (1992) ou em Paulino *et al.* (2003).

O método de Geweke verifica a convergência do MCMC quando se tem uma número N de observações suficientemente grande. Suponha que queremos estimar θ e para isso foi gerado um MCMC, para a averiguação da convergência utilizando Geweke, calcula-se as seguintes médias,

$$g_a = \frac{1}{n_a}\sum_j \theta^j \text{ e } g_b = \frac{1}{n_b}\sum_j \theta^j$$

sendo que j indica a posição na cadeia, g_a é o valor da média que foi calculada utilizando as n_a primeiras posições da cadeia e g_b é o valor da média que foi calculada utilizando as n_b últimas posições da cadeia.

Se a cadeia convergiu temos que g_a deve ser semelhante a g_b, admitindo que n_a/N e n_b/N são fixo e $N \longrightarrow \infty$; então pode se mostrar que

$$\frac{(g_a - g_b)}{\sqrt{(s_a^2/n_a) + (s_b^2/n_b)}} \longrightarrow N(0,1),$$

sendo que, s_a^2 e s_b^2 são estimativas independentes das variância assintóticas de g_a e g_b.

Assumimos que houve convergência se o valor desta estatística ficar entre $\pm$ 1,96. Este método está implementado em um pacote do R Development Core Team (2008) chamado CODA (Plummer *et al.*, 2009).