Modelos de Sobrevivência na Presença de Eventos Recorrentes e Longa Duração

Juliana Cobre

Orientador: Prof. Dr. Francisco Louzada Neto

São Carlos

Março de 2010

Modelos de Sobrevivência na Presença de Eventos Recorrentes e Longa Duração

Juliana Cobre

Orientador: Prof. Dr. Francisco Louzada Neto

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

São Carlos

Março de 2010

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária/UFSCar

C657ms

Cobre, Juliana.
Modelos de sobrevivência na presença de eventos recorrentes e longa duração / Juliana Cobre. -- São Carlos : UFSCar, 2010.
80 f.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2010.

1. Análise de sobrevivência. 2. Inferência clássica. 3. Inferência bayesiana. 4. Riscos competitivos. 5. Fração de cura. I. Título.

CDD: 519.9 (20ª)

Juliana Cobre

Modelos de sobrevivência na presença de eventos recorrentes e longa duração

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

Aprovada em 05 de março de 2010.

BANCA EXAMINADORA

Presidente _____
Prof. Dr. Francisco Louzada Neto (DEs-UFSCar/Orientador)

1° Examinador _____
Prof. Dr. Dani Gamerman (IM-UFRJ)

2° Examinador _____
Prof. Dr. Edwin Moises Marcos Ortega (ESALQ-USP)

3° Examinador _____
Prof. Dr. Heleno Bolfarine (IME-USP)

4° Examinador _____
Prof. Dr. Vicente Garibay Cancho (ICMC-USP)

Agradecimentos

A Deus pela saúde e pela força para superar as dificuldades. Por ter colocado em meu caminho pessoas importantes que de uma forma ou de outra muito me ajudaram.

Ao meu orientador Francisco Louzada Neto por ter confiado em mim para o desenvolvimento do projeto, pelo apoio e pelo direcionamento da pesquisa. Suas sugestões foram essenciais para a realização dos trabalhos.

Ao Professor e amigo Mário de Castro, pela contribuição científica, pelo incentivo à pesquisa e pela paciência em esclarecer minhas dúvidas. Foi fundamental seu auxílio na programação.

Aos Professores Dani Gamerman, Heleno Bolfarine, Edwin M. M. Ortega e Vicente G. Cancho, pelas sugestões e correções importantes à melhoria deste trabalho.

À Professora Gleici S. C. Perdoná por ter fornecido os dados e estabelecido o contato com a equipe médica da FMRP-USP.

Aos funcionários do Departamento de Estatística da UFSCar, especialmente à Isabel Araujo, pelos serviços gentilmente prestados.

Ao amigo Erlandson, por ter me iniciado na programação do sistema R. Ao amigo Roberto, por ter se disponibilizado a ajudar com as simulações, o que reduziu o tempo de espera dos resultados. À amiga Sandra, por ter me convencido a seguir em frente, pela companhia nas participações em congressos e nos momentos de diversão. À amiga Viviane, por ter me dado oportunidade de ministrar aulas, o que me trouxe realização pessoal e profissional. Amigos confidentes e incentivadores!

Agradeço especialmente aos meus pais, João e Edna, que sempre me estimularam a prosseguir meus estudos e me deram o conforto para cumpri-los. São fontes de força na minha luta em sempre dar o melhor de mim. Exemplos de honestidade e respeito.

E sem a companhia, o carinho e a compreensão do Marcel todo este período teria sido muito mais difícil.

Resumo

Neste trabalho propomos analisar dados de eventos recorrentes, dados de eventos recorrentes com fração de cura e dados de eventos recorrentes com tempos não observados e causas competitivas, que implicam na possibilidade de cura. Para a análise de dados de evento recorrente propomos um modelo de escala múltipla de tempo, que engloba diversas classes de modelos como casos particulares. Na análise de dados de eventos recorrentes com fração de cura tivemos como base os modelos de escala múltipla de tempo e o modelo de mistura padrão. Também propomos um modelo geral para tratar de dados na presença de causas competitivas. Neste caso, assumimos que o número de causas competitivas segue uma distribuição binomial negativa generalizada e consideramos duas abordagens para o tempo de ocorrência de cada causa, sendo uma delas uma distribuição Weibull e a outra uma distribuição log-logística. Para todos os modelos propostos foram feitos estudos de simulação com o objetivo de analisar as propriedades frequentistas dos processos de estimação. Aplicações a conjuntos de dados reais mostraram a aplicabilidade dos modelos propostos.

Abstract

In this thesis it is proposed to analyze recurrent event data, recurrent event data with cure fraction and recurrent event data with censoring and competing causes. For the recurrent event data analysis it is proposed a multiple time scale survival model, which includes several particular cases. For recurrent event data with a cure fraction we consider a multiple time scale survival models embedded on a mixture cure fraction modeling. It is also proposed a general model to survival data in presence of competitive causes. In this case, it is assumed that the number of competitive causes follows a generalized negative binomial distribution. While, for the time of occurrence of each cause, a Weibull and a log-logistic distribution were considered. Simulations studies were conducted for every proposed model in order to analyze the asymptotical properties of the estimation procedures. Both, maximum likelihood and Bayesian approaches were considered for parameter estimation. Real data applications demonstrate de use of the proposed models.

Sumário

Lista de Figuras

Lista de Tabelas

Introdução

O desenvolvimento e o aprimoramento de técnicas estatísticas, assim como o avanço computacional das últimas três décadas, implicaram no grande desenvolvimento da análise de sobrevivência, uma das áreas da Estatística que teve um crescimento significativo neste período (Colosimo & Giolo, 2006).

Em análise de sobrevivência e confiabilidade o foco do estudo é o tempo do acontecimento de um certo evento, chamado de tempo de falha ou tempo de ocorrência. O evento de interesse pode ser a morte de um paciente, o aparecimento de um tumor, a falha de um componente eletrônico, o uso do cartão de crédito, um abalo sísmico, entre outros. Com exceção do primeiro exemplo, os demais podem ocorrer diversas vezes para o mesmo indivíduo, ou seja, são eventos recorrentes. Ignorar a recorrência do evento pode comprometer a eficácia da metodologia estatística. Diversos procedimentos metodológicos estatísticos existentes na literatura acomodam tais conjuntos de dados.

Motivados pelos avanços dos tratamentos médicos, pesquisadores passaram a estudar a possibilidade de o indivíduo deixar de ser suscetível ao evento de interesse, o que resultou na análise de sobrevivência com fração de cura (ou de longa duração). Os modelos propostos neste trabalho são motivados por estudos clínicos em que a recorrência de eventos e a possibilidade de cura estão presentes. De forma resumida, podemos dizer que a proposta está dividida em três partes: análise de eventos recorrentes, análise de eventos recorrentes com fração de cura e análise de eventos recorrentes com tempos não observados, fração de cura e causas competitivas.

No Capítulo 1 propomos um modelo de sobrevivência de escala múltipla de tempo para analisar dados de eventos recorrentes. A estrutura do modelo acomoda uma ampla classe de modelos de sobrevivência, incluindo o modelo de Poisson, modelos de renovação e de contagem como casos particulares. Analisamos aplicações para um número pequeno ou moderado de indivíduos observados e para um número pequeno ou moderado de eventos

por indivíduo. Pela inferência bayesiana obtivemos as estimativas dos parâmetros. Estudos de simulação foram realizados para analisar as propriedades frequentistas do processo de estimação. O modelo proposto foi aplicado a um conjunto de dados reais. Os resultados obtidos neste capítulo foram condensados nos artigos Cobre & Louzada Neto (2009) e Louzada Neto & Cobre (2010).

Com o objetivo de analisar dados de eventos recorrentes com fração de cura, propomos no Capítulo 2 o modelo de escala múltipla de tempo com longa duração. A proposta é baseada em Boag (1949) e Berkson & Gage (1952) e no modelo de escala múltipla de tempo. Realizamos estudos de simulação para verificar as propriedades frequentistas do procedimento de estimação. Um conjunto de dados reais foi utilizado para ilustrar a aplicabilidade do modelo proposto. Deste capítulo proveio o relatório técnico Louzada Neto & Cobre (2008) submetido à publicação.

Em diversas situações a ocorrência de um evento pode se dar devido a uma dentre várias causas competitivas. Tanto o número de causas assim como o tempo de sobrevivência associado a cada causa não são observados. Se a probabilidade de o número de causas competitivas ser igual a zero for não nula temos a possibilidade de fração de cura. Nos Capítulos 3 e 4 supomos que o número de causas competitivas segue uma distribuição binomial negativa generalizada proposta por Hanin (2001). A principal vantagem desta suposição é estimar duas importantes taxas do processo: a taxa do aumento do número de eventos e a eficácia da intervenção para a diminuição da ocorrência do evento. Propomos duas possibilidades para a distribuição dos tempos de cada causa: no Capítulo 3 a distribuição Weibull e no Capítulo 4 a distribuição log-logística. A primeira delas acomoda funções de riscos crescentes, decrescentes e constantes. A distribuição log-logística ajusta funções de risco decrescentes ou que primeiramente crescem e depois decrescem. Os estudos de simulação foram realizados para verificar a possibilidade de estimação dos parâmetros, assim como as propriedades frequentistas das estimativas clássica e bayesiana dos parâmetros do modelo. Os modelos foram ajustados a dois conjuntos de dados reais para exemplificar a abordagem e a interpretação dos parâmetros. Resultaram do Capítulo 3 os relatórios técnicos Cobre, Louzada Neto & Perdoná (2009), submetido à publicação, e Cobre, Louzada Neto & Perdoná (2010). O relatório técnico Louzada Neto, Cobre & Perdoná (2009) é proveniente do Capítulo 4 e também foi submetido à publicação.

Capítulo 1

Modelo de Sobrevivência de Escala Múltipla de Tempo

Dados de sobrevivência em que o evento de interesse pode ocorrer mais de uma vez para o mesmo indivíduo existem em diversas áreas como, por exemplo, a biomédica, a criminológica, a demográfica, a industrial e a financeira. A observação do número de ocorrências para cada indivíduo, os indicadores de censura, assim como os instantes de suas ocorrências e possíveis covariáveis formam o conjunto de dados de cada indivíduo. Em geral o objetivo do estudo é explicar a natureza da variação entre os indivíduos em termos das covariáveis e das possíveis censuras dos dados.

Dentre os diversos modelos que visam descrever o processo está a representação de Poisson (Lawless, 2002), que modela o tempo total de estudo desconsiderando os tempos entre as ocorrências. Já o processo de renovação (Prentice, Willians & Peterson, 1981) modela os intervalos entre as ocorrências de tempo, implicando que o risco da próxima ocorrência não tem início antes de a anterior ter ocorrido. Um outro tipo de modelo é o processo de Poisson-renovação (Cox, 1972), em que ambas as escalas, tempo total e tempo intervalar (tempo entre duas ocorrências sucessivas), são consideradas no estudo. Wei, Lin & Weissefeld (1989) propõem modelos de risco proporcional para o tempo de vida, Pepe & Cai (1993) consideram funções de taxa, Nelson (1988, 1995) apresentam métodos gráficos, Lawless & Nadeau (1995) e Louzada Neto (2004) abordam modelos mais gerais com a inclusão de covariáveis, que consideram duas escalas de tempo, tempo total e tempo intervalar.

Neste trabalho propomos um modelo de escala múltipla de tempo para modelar da-

dos de eventos recorrentes. A ideia do modelo é combinar tempo total, tempo intervalar e a contagem de eventos em um modelo híbrido, ou seja, admitir múltiplas escalas de tempo além da contagem do número de eventos para cada indivíduo. O modelo proposto engloba uma ampla classe de modelos, incluindo o processo de Poisson e o processo de renovação como casos particulares. Na versão do modelo com covariáveis, assumimos uma função base proporcional. Analisamos situações em que um número moderado ou grande de indivíduos é observado e o número de eventos por indivíduos pode ser pequeno ou moderado. Os procedimentos de inferência foram dados através de reamostragem, sendo as amostras construídas com o método MCMC (Markov Chain Monte Carlo). Estudos de simulação foram realizados baseados em um cenário médico com o objetivo de verificar algumas propriedades frequentistas do procedimento de estimação para diferentes tamanhos amostrais.

Este capítulo está organizado como segue. O modelo proposto está descrito na Seção 1.1. Os procedimentos de inferência, a validação do modelo e os resultados dos estudos de simulação são apresentados na Seção 1.2. A análise de um conjunto de dados reais é apresentada na Seção 1.3. A Seção 1.4 finaliza o capítulo com alguns comentários finais.

1.1 Modelo

Os dados para o j-ésimo indivíduo consistem no número total, $m_j \geq 0, m_j \in \mathbb{N}$, de eventos observados no intervalo $(0, \tau_j]$ e os tempos das m_j ocorrências, $0 \leq t_{j1} < t_{j2} < \cdots < t_{jm_j}$, $j = 1, \ldots, n$. Primeiramente definimos a função de sobrevivência parcial por

$$S(t_{j_i}|\cdot) = \exp\left\{-H_i(t_{j_i}|\cdot)\right\},$$

em que H_i é uma função definida sobre o intervalo de tempo $[t_{j_i}, t_{j_{i+1}})$, $i = 0, \ldots, m_j$. Então a função de sobrevivência total é dada por

$$S(t_j|\cdot) = \exp\left\{-\sum_{i=0}^{m_j} H_i(t_{j_i}|\cdot)\right\}. \tag{1.1}$$

Propomos um modelo de sobrevivência de escala múltipla de tempo (EMT) que considera duas escalas de tempo, tempo intervalar e tempo total, e também o número de eventos para cada indivíduo assumindo que

$$H_i(t_{j_i}|\cdot) = q_1(x_{t_{j_i}}; \boldsymbol{\theta_1}) q_2(t_{m_j}; \boldsymbol{\theta_2}) q_3(i; \boldsymbol{\theta_3}) g(\boldsymbol{\beta}^T \boldsymbol{z}_j), \tag{1.2}$$

em que $q_1(\cdot), q_2(\cdot)$ e $q_3(\cdot)$ são funções positivas denotando a função intensidade-base paramétrica sobre o intervalo de tempo, $x_{t_{j_i}} = t_{j_i} - t_{j_{i-1}}$, tempo total, t_{j_i}, e contagem de eventos ocorridos em $(0, t_{j_i}]$, i, respectivamente, com vetores de parâmetros desconhecidos $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}$ e $\boldsymbol{\theta_3}$; $g(\cdot)$ é uma função conhecida, igual a um quando seu argumento é zero e $\boldsymbol{\beta}$ é um vetor de coeficientes de regressão. As covariáveis $\boldsymbol{z}_j$ são assumidas fixas, de modo que não são afetadas pelo processo.

O modelo (1.1) com (1.2) abrange uma ampla classe de modelos de sobrevivência e seus casos particulares podem ser obtidos com um número reduzido de parâmetros. Para obter o processo de Poisson não-homogêneo basta tomar $q_1(\cdot) = q_3(\cdot) = 1$ (Lawless, 1982). Considerando $q_2(\cdot) = q_3(\cdot) = 1$ temos o processo de renovação (Prentice *et al.*, 1981). E apenas $q_3(\cdot) = 1$, o processo de Poisson-renovação (Cox, 1972).

Analisaremos um modelo paramétrico considerando uma parametrização particular bastante flexível de (1.2), em que $q_1(x_t; \boldsymbol{\theta_1}) = q_1(x_t; \gamma) = \gamma x_t^{\gamma-1}$, $q_2(t; \boldsymbol{\theta_2}) = q_2(t; \phi) = 1 + \phi t$, $q_3(i; \boldsymbol{\theta_3}) = q_3(i; \psi) = \psi^{i-1}$ e $g(\boldsymbol{\beta}^T \boldsymbol{z}) = \exp(\boldsymbol{\beta}^T \boldsymbol{z})$. Dessa forma, de (1.1) e (1.2), a função sobrevivência parcial e a função sobrevivência global de nosso modelo EMT são dadas, respectivamente, por

$$S(t_{j_i}|\cdot) = \exp\left\{-\psi^{i-1} x_{j_i}^{\gamma} \left(1 + \phi t_{j_{i-1}} + \phi\gamma \frac{x_{j_i}}{\gamma+1}\right) \mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{z}_j}\right\} \tag{1.3}$$

e

$$S(t_j|\cdot) = \exp\left\{-\sum_{i=1}^{m_j} \psi^{i-1} x_{j_i}^{\gamma} \left(1 + \phi t_{j_{i-1}} + \phi\gamma \frac{x_{j_i}}{\gamma+1}\right) \mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{z}_j}\right\}, \tag{1.4}$$

em que ϕ, γ e ψ são parâmetros não-negativos, $\boldsymbol{\beta}$ é o vetor de parâmetros associado às covariáveis observadas, $\boldsymbol{z}_j$, $t_{j_1}, \ldots, t_{j_{m_j}}$ são os tempos de ocorrência do evento em estudo para o j-ésimo indivíduo, $x_{j_i} = t_{j_i} - t_{j_{i-1}}$ é o tempo intervalar entre ocorrências sucessivas com $t_{j_0} = 0$ e $\boldsymbol{\beta}^T \boldsymbol{z}_j = \beta_1 z_{j1} + \ldots + \beta_K z_{jK}$ não tem intercepto.

Uma vantagem dessa parametrização é sua facilidade de interpretação. Os parâmetros γ e ϕ denotam o efeito específico de cada escala de tempo (tempo total e tempo intervalar, respectivamente) na função sobrevivência, o parâmetro ψ denota o efeito do número de eventos na abordagem. Temos também que a componente de renovação, $q_1(\cdot)$, é dada por um modelo exponencial, enquanto que a componente de Poisson, $q_2(\cdot)$, funciona como um processo de Poisson dependente no tempo. A função de contagem de eventos, $q_3(\cdot)$, penaliza grande números de eventos se $\psi > 1$ e se $0 < \psi < 1$ o efeito é oposto. Além disso temos um efeito exponencialmente proporcional das covariáveis.

1.1.1 Modelos especiais

O modelo proposto abrange como casos particulares alguns modelos existentes na literatura, listados abaixo

- **Processo de contagem.** O modelo EMT com $\phi = 0$ e $\gamma = 1$ se reduz a

$$S(t_j|\cdot) = \exp\left\{-\sum_{i=1}^{m_j} \psi^{i-1} \mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{z}_j}\right\}, \tag{1.5}$$

  que é o processo de contagem (Cox, 1972).

- **Modelo de renovação Weibull ordinário.** Para $\phi = 0$ e $\psi = 1$, o modelo EMT se reduz a

$$S(t_j|\cdot) = \exp\left\{-\sum_{i=1}^{m_j} \psi^{i-1} x_{j_i}^{\gamma} \mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{z}_j},\right\} \tag{1.6}$$

  que é o modelo de renovação Weibull ordinário para os intervalos de tempo (veja p. ex. Yannaros, 1994).

- **Processo de Poisson não-homogêneo.** Se $\gamma = 1$ e $\psi = 1$, obtemos uma função de sobrevivência

$$S(t_j|\cdot) = \exp\left\{-\sum_{i=1}^{m_j} x_{j_i} \left(1 + \phi t_{j_i} + \phi \frac{x_{j_i}}{2}\right) \mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{z}_j}\right\}, \tag{1.7}$$

  que é o processo de Poisson não-homogêneo (Lawless, 1982, p. 494).

- **Modelo Weibull ordinário com contagem.** Para $\phi = 0$, a função do EMT se reduz a

$$S(t_j|\cdot) = \exp\left\{-\sum_{i=1}^{m_j} \psi^{i-1} x_{j_i}^{\gamma} \mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{z}_j}\right\}, \tag{1.8}$$

  que é um modelo de Weibull ordinário com parâmetro de contagem, então chamado de modelo Weibull ordinário com contagem (veja p.ex. McShane, Adrian, Bradlow & Fader, 2008).

- **Processo de Poisson-renovação.** Fixando $\psi = 1$ obtemos o processo de Poisson-renovação (Prentice *et al.*, 1981), cuja função parcial de sobrevivência é dada por

$$S(t_j|\cdot) = \exp\left\{-\sum_{i=1}^{m_j} x_{j_i}^{\gamma} \left(1 + \phi t_{j_i} + \phi\gamma \frac{x_{i_j}}{\gamma + 1}\right) \mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{z}_j}\right\}. \tag{1.9}$$

- **Processo de Poisson Não-homogêneo com contagem.** Se apenas $\gamma = 1$ obtemos o processo de Poisson não-homogêneo com parâmetro de contagem, ao qual chamamos de processo de Poisson não-homogêneo e contagem (veja p. ex. Massey, Parker & Whitt, 1996),

$$S(t_j|\cdot) = \exp\left\{-\sum_{i=1}^{m_j} \psi^{j-1} x_{j_i} \left(1 + \phi t_{j_i} + \phi\frac{x_{j_i}}{2}\right) \mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{z}_j}\right\}. \tag{1.10}$$

1.2 Inferência

Para a inferência adotamos um procedimento bayesiano. A função de verossimilhança, as distribuições *a priori* para os parâmetros do modelos, detalhes do algoritmo MCMC, comparação de modelos e estudos de simulação feitos para analisar as propriedades frequentistas do procedimento de estimação são descritos a seguir.

1.2.1 Função de verossimilhança

Sejam n indivíduos sujeitos à ocorrência de certo evento recorrente. Os dados referentes ao j-ésimo indivíduo são compostos por m_j, que denota o número total de eventos observados no período em estudo; $t_{j1} < t_{j2} < \ldots < t_{m_j}$ são os tempos de falha contínuos; $x_{j_i} = t_{j_i} - t_{j_{i-1}}$, com $t_{j_0} = 0$, denota os intervalos entre sucessivos eventos; $\boldsymbol{z_j}$ representa o vetor de covariáveis; e δ_j é a variável indicadora de censura tal que $\delta_j = 1$ se o tempo de falha é observado e $\delta_j = 0$, se o tempo de falha é censurado. Assim, o conjunto de dados observados, $\boldsymbol{D}$, é composto por três vetores $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n)$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ e $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, em que, para o j-ésimo indivíduo, $\boldsymbol{t}_j$ é o vetor de tempos de ocorrência do evento de interesse. A contribuição de cada indivíduo para a função de verossimilhança, $L_j(\phi, \gamma, \psi, \boldsymbol{\beta}|\boldsymbol{D})$, é dada pela função densidade se o indivíduo apresentou o evento de interesse e pela função de sobrevivência se o indivíduo foi censurado (Lawless, 2002), implicando que a função de verossimilhança para cada indivíduo no intervalo de tempo x_{j_i} é dada por

$$L_{j_i}(\cdot|\boldsymbol{D}_j) = f(t_{j_i}|\cdot)^{\delta_{j_i}} S(t_{j_i}|\cdot)^{1-\delta_{j_i}},$$

em que $\delta_{j_i} = 1$ para $i = 1, \ldots, m_j - 1$, já que $t_{j1} < t_{j2} < \ldots < t_{m_j-1}$ são os tempos de ocorrências para o j-ésimo indivíduo.

A função de verossimilhança é dada pelo produto da contribuição de cada indivíduo com relação a todas as ocorrências e a todos os intervalos. Portanto

$$L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\beta}|\boldsymbol{D}) = \prod_{j=1}^{n}\prod_{i=1}^{m_j} f(t_{j_i}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\beta})^{\delta_{j_i}} S(t_{j_i}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\beta})^{1-\delta_{j_i}}$$

e a função de log-verossimilhança é dada por

$$l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\beta}|\boldsymbol{D}) = \sum_{j=1}^{n}\sum_{i=1}^{m_j} \log\left\{ f(t_{j_i}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\beta})^{\delta_{j_i}} S(t_{j_i}|\phi, \gamma, \psi, \boldsymbol{\beta})^{1-\delta_{j_i}} \right\}.$$

Considerando o modelo EMT dado em (1.4) temos que a função de verossimilhança é dada por

$$\begin{aligned} l(\phi, \gamma, \psi, \boldsymbol{\beta}|\boldsymbol{D}) = \sum_{j=1}^{n}\sum_{i=1}^{m_j} & \delta_j\{\boldsymbol{\beta}^T\boldsymbol{z}_j + \log\gamma + (\gamma-1)\log x_{j_i} + \log(1+\phi t_{j_i}) + (i-1)\log\psi\} \\ & - \left\{1 + \phi\ t_{j_i} + \phi\frac{\gamma}{\gamma+1}x_{j_i}\right\} x_{j_i}^{\gamma}\psi^{i-1}\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{z}_j}. \end{aligned} \tag{1.11}$$

1.2.2 Inferência bayesiana

A inferência bayesiana é uma alternativa para a estimação dos parâmetros do modelo proposto, $(\phi, \gamma, \psi, \boldsymbol{\beta})$. Consideramos uma distribuição *a priori* conjunta própria para os parâmetros do modelo com o intuito de garantir que a distribuição *a posteriori* conjunta seja própria (Ibrahim, Chen & Sinha, 2001). A densidade *a priori* conjunta é dada por

$$\pi(\phi, \gamma, \psi, \boldsymbol{\beta}) = f_\Gamma(\phi|a_\phi, b_\phi) f_\Gamma(\gamma|a_\gamma, b_\gamma) f_\Gamma(\psi|a_\psi, b_\psi) \prod_{k=1}^{K} f_{\mathcal{N}}(\beta_k|0, \sigma^2_{\beta_k}), \tag{1.12}$$

em que $f_\Gamma(x|a,b) \propto x^{a-1}\mathrm{e}^{-bx}, x > 0$, ou seja, é a função densidade da distribuição gama com parâmetros de forma $a > 0$ e de escala $b > 0$, cuja média é a/b e cuja variância é igual a a/b^2; e $f_{\mathcal{N}}(\cdot|0,\sigma^2)$ é a função densidade de uma distribuição normal com média 0 e variância σ^2. Assumimos conhecidos os hiperparâmetros das distribuições *a priori* consideradas em (1.12). Combinando as funções verossimilhança e as densidades *a priori* obtemos a densidade *a posteriori*

$$\pi(\phi, \gamma, \psi, \boldsymbol{\beta}|\boldsymbol{D}) \propto \exp\{l(\phi, \gamma, \psi, \boldsymbol{\beta}|\boldsymbol{D})\}\, \pi(\phi, \gamma, \psi, \boldsymbol{\beta}), \tag{1.13}$$

em que $l(\phi, \gamma, \psi, \boldsymbol{\beta}|\boldsymbol{D})$ é dada por (1.11), $\pi(\phi, \gamma, \psi, \boldsymbol{\beta})$ por (1.12) e $\boldsymbol{D}$ é o conjunto de dados observados.

Integrando a densidade *a posteriori* dada em (1.13) com relação a cada um dos parâmetros obtemos as densidades marginais *a posteriori* de cada um dos parâmetros. Quando tais integrais não são analiticamente calculáveis, como é o nosso caso, uma alternativa é fazer uso de um dos métodos de Monte Carlo com cadeias de Markov (MCMC), como o amostrador de Gibbs e o algoritmo de Metropolis-Hastings (veja p. ex. Chib & Greenberg, 1995). Primeiramente devemos explicitar as densidades condicionais completas de todos os parâmetros, dadas por

$$\begin{aligned}\pi(\phi|\gamma,\psi,\boldsymbol{\beta},\boldsymbol{D}) \propto & \phi^{a_\phi-1}\exp\Bigg\{-b_\phi\phi+\sum_{j=1}^{n}\sum_{i=1}^{m_j}[\log(1+\phi)\\ & -(1+\phi t_{j_i}+\phi\frac{\gamma}{\gamma+1}x_{j_i})x_{j_i}^{\gamma}\psi^{i-1}\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{z}_j}]\Bigg\},\end{aligned}$$

$$\begin{aligned}\pi(\gamma|\phi,\psi,\boldsymbol{\beta},\boldsymbol{D}) \propto & \gamma^{a_\gamma-1}\exp\Bigg\{-b_\gamma\gamma+\sum_{j=1}^{n}\sum_{i=1}^{m_j}[\log\gamma+(\gamma-1)\log x_{j_i}\\ & -(1+\phi t_{j_i}+\phi\frac{\gamma}{\gamma+1}x_{j_i})x_{j_i}^{\gamma}\psi^{i-1}\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{z}_j}]\Bigg\}\end{aligned}$$

$$\begin{aligned}\pi(\psi|\phi,\gamma,\boldsymbol{\beta},\boldsymbol{D}) \propto & \psi^{a_\psi-1}\exp\Bigg\{-b_\psi\psi+\sum_{j=1}^{n}\sum_{i=1}^{m_j}[(i-1)\log\psi\\ & -(1+\phi t_{j_i}+\phi\frac{\gamma}{\gamma+1}x_{j_i})x_{j_i}^{\gamma}\psi^{i-1}\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{z}_j}]\Bigg\}\end{aligned}$$

e

$$\begin{aligned}\pi(\beta_k|\phi,\gamma,\psi,\boldsymbol{\beta}_{-k},\boldsymbol{D}) \propto & \exp\Bigg\{-\frac{1}{2\sigma_{\beta_k}^2}(\beta_k-\mu_{\beta_k})^2\\ & +\sum_{j=1}^{n}\sum_{i=1}^{m_j}[\boldsymbol{\beta}^T\boldsymbol{z}_j-(1+\phi t_{j_i}+\phi\frac{\gamma}{\gamma+1}x_{j_i})x_{j_i}^{\gamma}\psi^{i-1}\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{z}_j}]\Bigg\},\end{aligned}$$

em que a e b, indexados pelos parâmetros, são os parâmetros de forma e de escala da densidade gama das distribuições *a priori* de ϕ,γ e ψ; μ_{β_k} e σ_{β_k} são, respectivamente, as médias e os desvios padrão das distribuições *a priori* de cada um dos β_k; e $\boldsymbol{\beta}_{-k}=(\beta_1,\ldots,\beta_{k-1},\beta_{k+1},\ldots,\beta_K)$, ou seja, é o vetor de parâmetros $\boldsymbol{\beta}$ sem a k-ésima componente.

Devido às densidades condicionais dadas anteriormente não nos remeter a nenhuma distribuição conhecida, fazemos uso do algoritmo Metropolis-Hastings na geração dos valores de ϕ,γ,ψ e β_k. Tal algoritmo nos permite simular amostras de distribuições conjuntas complexas, utilizando as distribuições condicionais completas dos parâmetros desconhecidos, como mostra o esquema a seguir:

1. Inicie com $\boldsymbol{\theta}^{(0)}=(\phi^{(0)},\gamma^{(0)},\psi^{(0)},\boldsymbol{\beta}^{(0)})$.

2. Gere ϕ^* da distribuição *a priori* $\pi(\phi) = f_\Gamma(\phi|a_\phi, b_\phi)$ descrita anteriormente.

3. Gere um valor u da distribuição uniforme $U(0,1)$.

4. Se $u \leq \min\left\{1, \frac{\pi(\phi^*|\gamma^{(0)},\psi^{(0)},\boldsymbol{\beta}^{(0)},\boldsymbol{D}))}{\pi(\phi^{(0)}|\gamma^{(0)},\psi^{(0)},\boldsymbol{\beta}^{(0)},\boldsymbol{D})}\right\}$, então atualize $\phi^{(1)}$ por ϕ^*. Caso contrário, permaneça com $\phi^{(0)}$, ou seja, $\phi^{(1)} = \phi^{(0)}$.

5. Proceda analogamente para obter $\gamma^{(1)}$ e $\beta_k^{(1)}, k = 1, \ldots, K$.

6. Repita os passos de 2 a 5 até obter uma amostra de uma distribuição estacionária.

Para verificar a convergência do algoritmo Metropolis-Hastings, Gelfand & Smith (1990) sugerem o uso de técnicas gráficas, e Gelman & Rubin (1992) e Geweke (1992) propõem análises estatísticas dos dados da amostra gerada. O critério de Gelman-Rubin está implementado no sistema R (R Development Core Team, 2009) e será utilizado juntamente com a análise gráfica de Geweke.

1.2.3 Comparação de modelos

Devemos ser cautelosos na escolha de um modelo estatístico para representar os dados. O modelo EMT engloba diversos casos particulares e devemos verificar se um modelo mais simples pode ser considerado. Então devemos testar as hipóteses $H_0 : \psi = 1$, $H_0 : \phi = 0$, $H_0 : \gamma = 1$, $H_0 : \gamma = 1, \psi = 1$, $H_0 : \phi = 0, \psi = 1$ e $H_0 : \phi = 0, \gamma = 1$, que explicitam os diversos casos particulares de (1.4). Na literatura existem diversas metodologias que se propõem a analisar a adequabilidade do modelo, além de, dentre uma coleção deles, selecionar o melhor. Dentre as várias técnicas existentes (veja p. ex. Paulino, Turkman & Murteira, 2003, p. 348), podemos destacar o AIC (*Akaike Information Criterion*) e o BIC (*Bayesian Information Criterion*), que são definidos, respectivamente, por $-2l(\widehat{\boldsymbol{\theta}}_\rho) + 2g$ e $-2l(\widehat{\boldsymbol{\theta}}_\rho) + g\log(n)$, em que $\widehat{\boldsymbol{\theta}}_\rho$ é a estimativa de máxima verossimilhança sobre o modelo ρ, g é o número de parâmetros estimados sobre o modelo ρ, e n é a quantidade de observações na amostra. Melhores modelos correspondem a menores valores de AIC e BIC. Também podemos destacar o DIC (*Deviance Information Criterion*) proposto por Spiegelhalter, Best, Carlin & van der Linde (2002). Trata-se de uma aproximação de um dos primeiros critérios propostos na literatura, o fator de Bayes, e tem como objetivo incorporar a complexidade do modelo no critério de seleção, ou seja, de certa forma "penalizar"a verossimilhança. Seja $DL_\rho(\boldsymbol{\theta}_\rho) = -2\log\frac{f(\boldsymbol{D}|\theta_\rho,M_\rho)}{h(\boldsymbol{D})}$, em que $h(\boldsymbol{D})$ é uma

função dos dados que não interfere na escolha do modelo M_ρ, com vetor de parâmetros $\boldsymbol{\theta}_\rho$. Como medida de adequabilidade do modelo é proposto o valor esperado *a posteriori* de $DL_\rho(\boldsymbol{\theta}_\rho)$ e, associado à complexidade do modelo, é proposto um fator de penalização, p_{DL_ρ}, dado por

$$p_{DL_\rho} = \mathrm{E}_{(\boldsymbol{\theta}_\rho|\boldsymbol{D},M_\rho)}[DL_\rho(\boldsymbol{\theta}_\rho)] - DL_\rho[\mathrm{E}_{(\boldsymbol{\theta}_\rho|\boldsymbol{D},M_\rho)}(\boldsymbol{\theta}_\rho)].$$

O fator DIC é então dado por

$$DIC_\rho = \mathrm{E}_{(\boldsymbol{\theta}_\rho|\boldsymbol{D},M_\rho)}[DL_\rho(\boldsymbol{\theta}_\rho)] + p_{DL_\rho} = 2\mathrm{E}_{(\boldsymbol{\theta}_\rho|\boldsymbol{D},M_\rho)}[DL_\rho(\boldsymbol{\theta}_\rho)] - DL_\rho[\mathrm{E}_{(\boldsymbol{\theta}_\rho|\boldsymbol{D},M_\rho)}(\boldsymbol{\theta}_\rho)].$$

Menores valores do DIC indicam os modelos a serem escolhidos.

Em muitos casos, como é o nosso, os valores exigidos para o cálculo do DIC não são obtidos de forma analítica. No entanto, podemos obter aproximações numéricas recorrendo a métodos computacionais como por exemplo o MCMC. Neste caso, para calcularmos $\mathrm{E}_{(\boldsymbol{\theta}_\rho|\boldsymbol{D},M_\rho)}[DL_\rho(\boldsymbol{\theta}_\rho)]$, basta tomarmos a amostra da distribuição *a posteriori* $\boldsymbol{\theta}_\rho^* = \left\{\boldsymbol{\theta}_\rho^{(1)}, \ldots, \boldsymbol{\theta}_\rho^{(L)}\right\}$ utilizada para obter as estimativas dos parâmetros, veja algoritmo descrito na Seção 1.2.2. Com isso temos que

$$\mathrm{E}_{(\boldsymbol{\theta}_\rho|\boldsymbol{D},M_\rho)}[DL_\rho(\boldsymbol{\theta}_\rho)] \approx \frac{1}{L}\sum_{l=1}^{L} DL_\rho\left(\boldsymbol{\theta}_\rho^{(l)}\right).$$

Analogamente temos

$$\mathrm{E}_{(\boldsymbol{\theta}_\rho|\boldsymbol{D},M_\rho)}[\boldsymbol{\theta}_\rho] \approx \frac{1}{L}\sum_{j=1}^{L} \boldsymbol{\theta}_\rho^{(l)}.$$

1.2.4 Probabilidade de cobertura

Estudos de simulação foram realizados com o objetivo de analisar as propriedades frequentistas dos procedimentos de estimação. Para examinar as propriedades frequentistas construímos os intervalos de credibilidade para todos os parâmetros e calculamos suas probabilidades de cobertura (PC). Os valores dos parâmetros foram escolhidos baseados em um experimento clínico, em que é analisada a eficácia de um tratamento, veja Seção 1.3. O vetor de parâmetros a ser estimado é dado por $\boldsymbol{\lambda} = (\phi, \gamma, \psi, \beta)$. Consideramos $\phi = 0,8$, $\gamma = 2,0$, $\psi = 1,8$ e $\beta = -1,2$, diferentes tamanhos amostrais, $n = 30$, 50, 70 e 100, e diferentes números de ocorrência, $m = 3$ e 5, do evento de interesse. Supomos que em cada grupo há $n/2$ observações e para diferenciar os indivíduos pertencentes ao grupo

Tabela 1.1: Probabilidades de cobertura dos intervalos de credibilidade de 95% para cada parâmetro e para cada tamanho amostral com $m = 3$ ocorrências para cada observação.

Parâmetro	Tamanho amostral			
	30	50	70	100
ϕ	0,817	0,856	0,883	0,914
γ	0,862	0,834	0,864	0,931
ψ	0,877	0,885	0,887	0,862
β	0,957	0,952	0,956	0,982

em tratamento dos pertencentes ao grupo em controle, utilizamos uma covariável binária z, tal que z é igual a -1 ou 1. A distribuição $\Gamma(0,9; 0,3)$, cuja média é 3 e a variância é 10, foi considerada como distribuição *a priori* dos parâmetros ϕ, γ e ψ. Para o parâmetro β foi considerada uma distribuição normal com média 0 e variância 100. Foram gerados 1000 conjuntos de dados. Para cada conjunto de dados fictícios foram geradas duas cadeias de 5500 iterações, sendo que as 500 primeiras foram descartadas para eliminar a influência dos valores iniciais, e as restantes selecionadas de 20 em 20 para reduzir a correlação nas séries, resultando em uma amostra com 500 valores. Utilizamos o sistema R (R Development Core Team, 2009) em todo o estudo. Monitoramos a convergência das cadeias usando o método proposto por Gelman & Rubin (1992) e a análise gráfica proposta por Geweke (1992).

A fim de obtermos a PC dos intervalos de credibilidade, para todas as amostras calculamos os intervalos de credibilidade de 95% e verificamos se continham os respectivos verdadeiros valores dos parâmetros. Os resultados da PC para diferentes tamanhos amostrais e diferentes números de recorrências estão organizados nas Tabelas 1.1 e 1.2. A Tabela 1.1 nos permite concluir que para um número pequeno de recorrências, menos do que 100 observações podem comprometer as estimativas de ϕ, γ e ψ. Se o número de ocorrências subir para 5, a Tabela 1.2 nos mostra que números pequeno e moderado de observações não prejudicam as estimativas.

Tabela 1.2: Probabilidades de cobertura dos intervalos de credibilidade de 95% para cada parâmetro e para cada tamanho amostral com $m = 5$ ocorrências para cada observação.

Parâmetro	Tamanho amostral			
	30	50	70	100
ϕ	0,921	0,909	0,867	0,900
γ	0,972	0,977	0,983	0,988
ψ	0,910	0,940	0,932	0,956
β	0,944	0,947	0,955	0,954

1.3 Aplicação

Utilizamos os dados apresentados na Tabela 1 de Gail, Santner & Brown (1980), que apresentam os tempos de aparecimento de tumores mamários de 48 ratas em um experimento oncológico. Aleatoriamente 23 ratas foram selecionadas para receber certo tratamento, as 25 restantes formaram o grupo de controle. Todo o grupo foi induzido a não apresentar tumores durante os primeiros 60 dias. Posteriormente foram observadas por mais 122 dias, recebendo ou não tratamento dependendo de seu grupo. O número médio de recorrências no grupo em tratamento é 3,5 e no grupo em controle é 5,7. Ajustamos o modelo EMT e seus casos particulares a estes dados.

As distribuições *a priori* para os parâmetros utilizadas são $\phi \sim \Gamma(1, 4)$, $\gamma \sim \Gamma(1, 4)$, $\psi \sim \Gamma(1, 4)$ e $\beta_k \sim \mathcal{N}(0, 25)$. Os valores dos hiperparâmetros foram escolhidos subjetivamente, mas tendo em mente um balanço entre assegurar a não informação e a convergência do algoritmo Metropolis-Hastings.

Foram geradas duas cadeias com 55000 iterações cada. As 5000 primeiras foram descartadas para minimizar a possível influência do ponto inicial da cadeia. As restantes foram selecionadas de 50 em 50 para evitar a correlação nas séries, resultando em uma amostra com 2000 valores. Todo o estudo foi implementado no sistema R. A convergência das cadeias foi verificada pelos métodos descritos na Seção 1.2.2. A Tabela 1.3 fornece as médias *a posteriori* e seus respectivos desvios padrão (entre parênteses). Os valores dos critérios AIC, BIC e DIC são apresentados na Tabela 1.4. Os resultados fornecem evidência a favor do modelo completo, mostrando a importância de serem levadas em consideração, na análise, a contagem de eventos assim como as duas escalas de tempo.

Tabela 1.3: Médias *a posteriori* e respectivos desvios padrão (entre parênteses).

Modelo	Parâmetro			
	γ	ϕ	ψ	β
EMT	0,246	0,095	0,762	0,869
	(0,105)	(0,102)	(0,135)	(0,570)
Modelo Weibull	0,235	-	0,764	0,790
ordinário com contagem	(0,091)	-	(0,124)	(0,545)
Processo de	0,336	0,102	-	0,145
Poisson-renovação	(0,083)	(0,402)	-	(0,036)
Processo não-homogêneo	-	0,090	0,776	0,652
com contagem	-	(0,085)	(0,113)	(0,552)
Processo de contagem	-	-	0,766	0,665
	-	-	(0,115)	(0,337)
Modelo Weibull ordinário	0,342	-	-	0,092
	(0,091)	-	-	(0,388)
Processo de Poisson	-	0,090	-	0,843
não-homogêneo	-	(0,086)	-	(0,468)

A estimativa de β nos dá evidência de que o tratamento é benéfico, o que está em concordância com as conclusões obtidas por Gail *et al.* (1980) e Lawless (1995). As Figuras 1.1 e 1.2 mostram os históricos das cadeias e as densidades *a posteriori* marginais.

1.4 Comentários finais

O modelo proposto EMT permite duas escalas de tempo (tempo intervalar e tempo total), a contagem de eventos e a inclusão de covariáveis. O modelo abrange diversos casos particulares que podem ser testados diretamente do modelo. As estimativas dos parâmetros do modelo foram obtidas através da inferência bayesiana, permitindo a incorporação de conhecimentos prévios e menor esforço computacional. Os resultados obtidos com o conjunto de dados reais mostram a importância da contagem de eventos e das duas escalas de tempo em sua análise.

Tabela 1.4: Valores dos critérios AIC, BIC e DIC.

Modelo	AIC	BIC	DIC
EMT	220,856	222,513	395,462
Processo de Poisson não-homogêneo	221,335	222,910	395,787
Modelo Weibull ordinário com contagem	230,562	246,298	397,999
Processo de Poisson-renovação	280,361	295,007	400,740
Processo não-homogêneo com contagem	296,091	301,632	462,446
Processo de contagem	320,551	340,196	472,270
Modelo Weibull ordinário	330,396	350,672	495,972

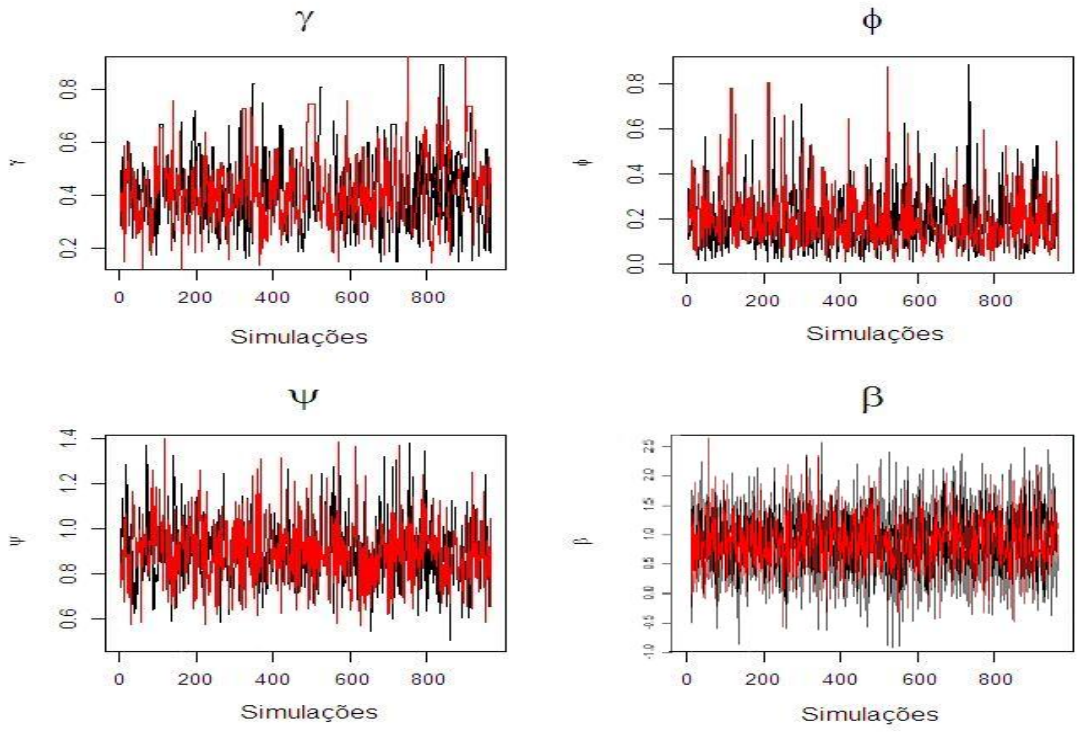

Figura 1.1: Histórico das cadeias.

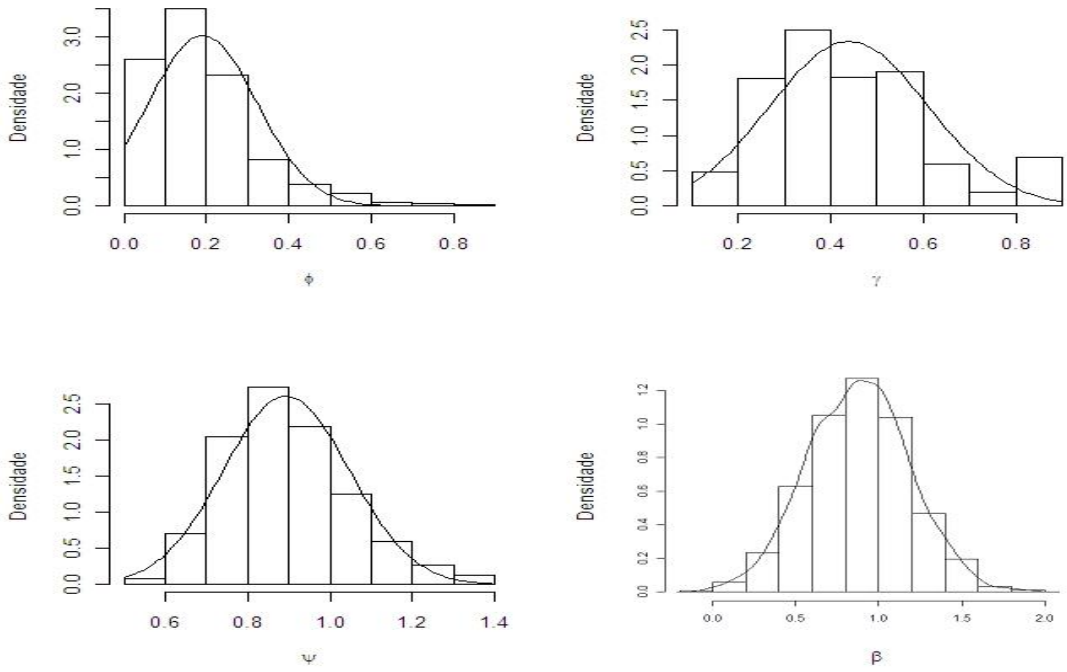

Figura 1.2: Densidades marginais *a posteriori*.

Capítulo 2

Modelo de Sobrevivência de Escala Múltipla de Tempo com Longa Duração

Modelos que consideram que uma parte da população pode ser ou se tornar não suscetível a certo evento de interesse têm sido amplamente desenvolvidos nos últimos tempos. Evidência de que uma fração de cura existe na população em estudo pode ser dada pela grande porcentagem de censuras no conjunto de dados (Goldman, 1984). Por exemplo, em estudos clínicos a população pode responder favoravelmente a um tratamento e ser considerada curada. Em estudos tradicionais, indivíduos que não apresentam o evento de interesse até certo momento são considerados censurados. Entretanto, não levam em consideração se os indivíduos não sofrerão tal evento. Devido a isso, esses modelos são chamados de modelos com fração de cura ou modelos de longa duração. Modelos de longa duração foram primeiramente abordados por Boag (1949) e Berkson & Gage (1952), que supõem a existência de apenas uma possível causa interferindo para a ocorrência do evento, e que esta causa se manifesta ou não segundo uma probabilidade a ser estimada. Mais recentemente, trabalhos como Yakovlev & Tsodikov (1996), Ibrahim *et al.* (2001) e Yin & Ibrahim (2005) propõem modelos mais abrangentes, no sentido de que permitem que não apenas uma, mas várias causas estejam relacionadas à ocorrência do evento de interesse.

Em muitos casos o evento de interesse pode ocorrer mais de uma vez para o mesmo indivíduo. Dados de sobrevivência com eventos recorrentes ocorrem em diversas áreas, como por exemplo, em estudos oncológicos com o reaparecimento de um tumor cancerígeno, em

estudos geológicos com a recorrência de um abalo sísmico, ou em estudos industriais com mais de uma falha de um componente. Uma das principais referências na modelagem de eventos recorrentes é o trabalho de Cox (1972), em que são considerados o tempo total de estudo e o tempo entre duas recorrências consecutivas, chamado de tempo intervalar. O trabalho de Prentice *et al.* (1981) considera o tempo intervalar implicando que o risco do próximo evento não tem início antes de o anterior ter ocorrido. A representação de Poisson apresentada em Lawless (1982) modela o tempo total de estudo, desconsiderando os tempos entre as ocorrências. Já em Louzada Neto (2004) e Louzada Neto (2008), ambas as escalas de tempo são consideradas, intervalar e tempo total, e também a contagem do número de eventos para cada indivíduo.

McDonald & Rosina (2001) e Yu (2008) propõem modelos de mistura para analisar situações em que tanto a recorrência do evento como a possibilidade de tornar-se fora de risco podem ser observadas. Yu (2008) analisa a possibilidade de longa duração num estudo de câncer colorretal após certa intervenção cirúrgica. Os pacientes que não desenvolveram novos tumores após cinco anos são considerados fora de risco, e pacientes nos quais a doença reincide fazem parte do grupo em risco.

Neste trabalho propomos um modelo de escala múltipla de tempo combinado com o trabalho de Berkson & Gage (1952) para a modelagem de eventos recorrentes com longa duração. O principal objetivo é analisar a eficácia de certa intervenção para impedir a ocorrência do evento de interesse na presença de covariáveis e das censuras. O modelo geral acomoda uma ampla classe de modelos incluindo o modelo de Poisson, o processo de renovação e modelos de contagem como casos especiais. A ideia é combinar as duas escalas de tempo, tempo total e tempo intervalar (intervalo entre eventos sucessivos), e a contagem de eventos, e decidir sobre sua importância no ajuste aos conjuntos de dados. Analisamos aplicações em que um número pequeno ou moderado de indivíduos é observado e o número de eventos por indivíduo pode ser pequeno.

O modelo proposto é descrito na Seção 2.1. Procedimentos de inferência, seleção de modelos e os resultados de simulações são apresentados na Seção 2.2. Um conjunto de dados reais foi analisado na Seção 2.3. Comentários finais na Seção 2.4 concluem o capítulo.

2.1 Modelo

Em estudos tradicionais envolvendo dados de sobrevivência, os indivíduos que não apresentam o evento de interesse até certo momento são censurados, sem se saber se o indivíduo deixou de ser suscetível a tal evento. Os indivíduos que, por diferentes motivos, abandonam o estudo também são censurados. Logo a censura pode significar tanto que o indivíduo deixou de ser suscetível a certo evento, como que o indivíduo abandonou o estudo. Na área médica ou biológica, um dos motivos do abandono do estudo é o falecimento do indivíduo. Portanto, é viável considerar um modelo que permita descrever tal heterogeneidade, como é o caso dos modelos de mistura de duas distribuições para o tempo em estudo. Nesta modelagem uma distribuição representa os tempos de falha ou sobrevivência dos indivíduos suscetíveis ao evento (em risco - ER) e a outra representa os tempos de sobrevivência dos indivíduos não-suscetíveis ao evento (fora de risco - FR). Esta última distribuição deve permitir tempos de sobrevivência infinitos, ou seja, deve ser degenerada (Maller & Zhou, 1996). O termo "longa duração" refere-se aos indivíduos não-suscetíveis ao evento de interesse. Na área médica, é comum se usar o termo "curado" para se referir à parte da população que não está mais em risco. Apenas para simplificar a linguagem, usaremos o termo curado.

Primeiramente, os modelos de mistura foram abordados por Boag (1949) e por Berkson & Gage (1952). Ambos os trabalhos consideram que o indivíduo pertence ou não ao grupo em risco com certa probabilidade. Matematicamente, a ideia é descrita a seguir. Seja T uma variável aleatória que representa o tempo até a ocorrência do evento de interesse, e θ a probabilidade de o indivíduo pertencer ao grupo FR. Considerando uma população em que existe a possibilidade de cura, a probabilidade de o evento ocorrer após o tempo t, para um indivíduo qualquer, é dada por

$$\begin{aligned} P(T > t) &= P(T > t|\text{FR})P(\text{FR}) + P(T > t|\text{ER})P(\text{ER}) \\ &= \theta P(T > t|\text{FR}) + (1-\theta)P(T > t|\text{ER}). \end{aligned}$$

Em termos da função de sobrevivência podemos escrever a expressão anterior como

$$S_{\text{pop}}(t|\cdot) = \theta S_{\text{FR}}(t|\cdot) + (1-\theta)S(t|\cdot), \tag{2.1}$$

em que $S_{\text{FR}}(t|\cdot)$ e $S(t|\cdot)$ são as funções de sobrevivência, respectivamente, dos indivíduos FR e ER, e S_{pop} denota a função de sobrevivência da população.

Os indivíduos FR não apresentarão o evento de interesse, ou seja, seu tempo de falha é infinito, o que nos dá

$$S_{\mathrm{FR}}(t|\cdot) = P(T > t|\mathrm{FR}) = 1, \quad \forall t \geq 0.$$

A função em (2.1) é então dada por

$$S_{\mathrm{pop}}(t|\cdot) = \theta + (1-\theta)S(t|\cdot).$$

Já os indivíduos em risco em algum momento apresentarão o evento de interesse, ou seja,

$$\lim_{t\to\infty} S(t|\cdot) = 0.$$

Consequentemente, temos

$$\lim_{t\to\infty} S_{\mathrm{pop}}(t|\cdot) = \theta,$$

e portanto a função de sobrevivência (não condicional) é imprópria e seu limite corresponde à proporção dos indivíduos fora de risco.

Na literatura existem várias propostas para a distribuição do tempo de sobrevivência dos indivíduos em risco. A família Weibull foi abordada por Farewell (1982), Ghitany & Maller (1992), Ghitany, Maller & Zhou (1994), Ng, McLachlan, Yau & Lee (2004). Peng, Dear & Denham (1998) propõem o uso da distribuição F generalizada. Na modelagem de eventos recorrentes podemos citar o processo de Poisson não-homogêneo e o processo de renovação puro que depende do tempo total de estudo (Lawless, 2002, p. 532). O processo de renovação (Prentice *et al.*, 1981) modela o tempo entre as diversas ocorrências e a última. O processo semi-markoviano considera cada ocorrência como sendo um estrato e que o indivíduo permanece no estrato até que a próxima ocorrência ou censura ocorra (Prentice *et al.*, 1981). Outra classe de modelos é composta pelo processo renovação de Poisson (Cox, 1972) e os modelos de escala híbrida (Louzada Neto, 2004, 2008), que incorporam duas escalas de tempo, tempo total e tempo intervalar. Modelos semiparamétricos foram abordados por Kuk & Chen (1992), Sy & Taylor (2000) e Yu (2008). Propomos um modelo de sobrevivência de escala múltipla de tempo para descrever a distribuição do tempo de sobrevivência dos indivíduos ER. O modelo tem como vantagem incorporar duas escalas de tempo, intervalar e tempo total, além da contagem do número de eventos.

2.1.1 Modelo de sobrevivência de escala múltipla de tempo

Os dados para o j-ésimo indivíduo são compostos pelo número total de eventos, m_j, observados no período de tempo $(0, \tau_j]$ e os tempos de ocorrência $t_{j_0} < t_{j_1} < \ldots < t_{j_{m_j}} \leq \tau_j$. No Capítulo 1 definimos a função parcial de sobrevivência por

$$S(t_{j_i}|\cdot) = \exp\left\{-H_i(t_{j_i}|\cdot)\right\}, \tag{2.2}$$

em que H_i é a função de risco definida em $[t_{j_{i-1}}, t_{j_i})$. Então, a função de sobrevivência completa é dada por

$$S(t_j|\cdot) = \exp\left\{-\sum_{i=1}^{m_j} H_i(t_{j_i}|\cdot)\right\}. \tag{2.3}$$

Seguindo Lawless & Thiagarajah (1996), que consideram um processo de Poisson e renovação abrangendo ambas as escalas de tempo, com o objetivo de estimação de um modelo específico, propomos um modelo de sobrevivência de escala múltipla de tempo (EMT), que inclui tanto o tempo intervalar quanto o tempo total de estudo, e também o número de eventos para cada indivíduo, assumindo que

$$H_i(t_{j_i}|\boldsymbol{\lambda}^*) = q_1(x_{t_{j_i}}; \boldsymbol{\theta_1}) q_2(t_{m_j}; \boldsymbol{\theta_2}) q_3(i; \boldsymbol{\theta_3}). \tag{2.4}$$

em que $q_1(x_t; \boldsymbol{\theta_1}) = q_1(x_t; \gamma) = \gamma x_t^{\gamma-1}$, $q_2(t; \boldsymbol{\theta_2}) = q_2(t; \phi) = 1 + \phi t$, $q_3(i; \boldsymbol{\theta_3}) = q_3(i; \psi) = \psi^{i-1}$ e $\boldsymbol{\lambda}^* = (\phi, \gamma, \psi)$. De (2.2) e (2.3), sua função de sobrevivência parcial e função sobrevivência são dadas, respectivamente, por

$$S(t_{j_i}|\boldsymbol{\lambda}^*) = \exp\left\{-\psi^{i-1} x_{j_i}^{\gamma} \left(1 + \phi t_{j_i} + \phi\gamma \frac{x_{j_i}}{\gamma + 1}\right)\right\}$$

e

$$S(t_j|\boldsymbol{\lambda}^*) = \exp\left\{-\sum_{i=1}^{m_j} \psi^{i-1} x_{j_i}^{\gamma} \left(1 + \phi t_{j_i} + \phi\gamma \frac{x_{j_i}}{\gamma + 1}\right)\right\}, \tag{2.5}$$

em que ϕ, γ e ψ são parâmetros não-negativos, $t_{j_0}, t_{j_1}, \ldots, t_{j_{m_j}}$ são os tempos de recorrência do evento em estudo para o j-ésimo indivíduo e $x_{j_i} = t_{j_i} - t_{j_{i-1}}$ o intervalo de tempo entre duas ocorrências sucessivas com $t_{j_0} = 0$. Essa formulação considera que a componente de renovação (tempo intervalar) é modelada por uma distribuição exponencial com parâmetro γ, a componente de Poisson (tempo total) modelada por um processo de Poisson dependente do tempo com parâmetro ϕ, e a contagem de evento penalizando grandes números de eventos se $\psi > 1$.

2.1.2 EMT com longa duração

Seguindo Berkson & Gage (1952), de (2.3) o modelo EMT com longa duração é dado por

$$S_{\text{pop}}(t|\boldsymbol{\lambda}^*, \theta) = \theta + (1-\theta)S(t|\boldsymbol{\lambda}^*). \tag{2.6}$$

Vale ressaltar que se $\theta = 0$, então $S_{\text{pop}}(t|\boldsymbol{\lambda}^*, \theta) = S(t|\boldsymbol{\lambda}^*)$, ou seja, (2.6) engloba a análise de sobrevivência usual. Uma de suas principais vantagens está na facilidade de inclusão de covariáveis (Yamaguchi, 1992). Descrevemos a proporção de curados da população em termos das covariáveis por uma função de ligação logística,

$$\theta_j = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{z}_j)}{1 + \exp(\boldsymbol{\beta}^T \boldsymbol{z}_j)}, \tag{2.7}$$

Dessa forma a probabilidade de cura é diferente para cada indivíduo.

Assumindo que $S(t_j|\boldsymbol{\lambda}^*)$ é dada por (2.5) e sendo $f(t_j|\boldsymbol{\lambda}^*) = -\frac{d}{dt}S(t|\boldsymbol{\lambda}^*)|_{t=t_j}$, a função de sobrevivência imprópria e a função densidade imprópria para o j-ésimo indivíduo do modelo EMT com longa duração são dadas, respectivamente, por

$$S_{\text{pop}}(t_j|\boldsymbol{\lambda}^*, \theta_j) = \theta_j + (1-\theta_j)S(t_j|\boldsymbol{\lambda}^*) \tag{2.8}$$

e

$$f_{\text{pop}}(t_j|\boldsymbol{\lambda}^*, \theta_j) = f(t_j|\boldsymbol{\lambda}^*)(1-\theta_j).$$

2.1.3 Modelos especiais

O modelo proposto abrange como casos particulares alguns modelos listados abaixo.

- **Processo de contagem com longa duração.** O modelo EMTLD com $\phi = 0$ e $\gamma = 1$ se reduz a

$$S_{\text{pop}}(t_j|\cdot, \boldsymbol{\theta}_j) = \boldsymbol{\theta}_j + (1-\boldsymbol{\theta}_j)\exp\left\{-\sum_{i=1}^{m_j}\psi^{i-1}\right\},$$

  que é o processo de contagem com longa duração.

- **Modelo de renovação Weibull ordinário com longa duração.** Para $\phi = 0$ e $\psi = 1$, o modelo EMT se reduz a

$$S_{\text{pop}}(t_j|\cdot, \boldsymbol{\theta}_j) = \boldsymbol{\theta}_j + (1-\boldsymbol{\theta}_j)\exp\left\{-\sum_{i=1}^{m_j}\psi^{i-1}x_{j_i}^{\gamma},\right\}$$

  que é o modelo de renovação Weibull ordinário para os intervalos de tempo com longa duração.

- **Processo de Poisson não-homogêneo com longa duração.** Se $\gamma = 1$ e $\psi = 1$, obtemos uma função de sobrevivência

$$S_{\text{pop}}(t_j|\cdot, \boldsymbol{\theta}_j) = \boldsymbol{\theta}_j + (1 - \boldsymbol{\theta}_j) \exp\left\{ -\sum_{i=1}^{m_j} x_{j_i} \left(1 + \phi t_{j_i} + \phi \frac{x_{j_i}}{2}\right) \right\},$$

que é o processo de Poisson não-homogêneo com longa duração.

- **Modelo Weibull ordinário com contagem com longa duração.** Para $\phi = 0$, a função do EMTLD se reduz a

$$S_{\text{pop}}(t_j|\cdot, \boldsymbol{\theta}_j) = \boldsymbol{\theta}_j + (1 - \boldsymbol{\theta}_j) \exp\left\{ -\sum_{i=1}^{m_j} \psi^{i-1} x_{j_i}^{\gamma} \right\},$$

que é um modelo de Weibull ordinário com parâmetro de contagem, então chamado de modelo Weibull ordinário com contagem com longa duração.

- **Processo de Poisson-renovação com longa duração.** Fixando $\psi = 1$ obtemos o processo de Poisson-renovação com longa duração, cuja função parcial de sobrevivência é dada por

$$S_{\text{pop}}(t_j|\cdot, \boldsymbol{\theta}_j) = \boldsymbol{\theta}_j + (1 - \boldsymbol{\theta}_j) \exp\left\{ -\sum_{i=1}^{m_j} x_{j_i}^{\gamma} \left(1 + \phi t_{j_i} + \phi\gamma \frac{x_{i_j}}{\gamma + 1}\right) \right\}.$$

- **Processo de Poisson não-homogêneo com contagem e longa duração.** Se apenas $\gamma = 1$ obtemos o processo de Poisson não-homogêneo com parâmetro de contagem, ao qual chamamos de processo de Poisson não-homogêneo com contagem e longa duração,

$$S_{\text{pop}}(t_j|\cdot, \boldsymbol{\theta}_j) = \boldsymbol{\theta}_j + (1 - \boldsymbol{\theta}_j) \exp\left\{ -\sum_{i=1}^{m_j} \psi^{j-1} x_{j_i} \left(1 + \phi t_{j_i} + \phi \frac{x_{j_i}}{2}\right) \right\}.$$

2.2 Inferência

Para a inferência adotamos o mesmo método bayesiano descrito na Seção 1.2.2. Para a comparação de modelos utilizamos os mesmos procedimentos descritos na Seção 1.2.3. A função de verossimilhança do modelo EMT com longa duração, as distribuições *a priori* dos parâmetros do modelo assim como a distribuição *a posteriori* são descritas a seguir.

2.2.1 Função de verossimilhança

Considerando que n indivíduos foram observados, o conjunto de dados é dado pelos vetores $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n)$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ e $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, sendo $\boldsymbol{t}_j$ o vetor dos tempos de ocorrência do evento para o j-ésimo indivíduo; δ_j a variável indicadora de censura do j-ésimo indivíduo, sendo $\delta_j = 0$ se o indivíduo for censurado e $\delta_i = 1$ caso contrário; e $\boldsymbol{z}_j$ o conjunto de covariáveis para cada indivíduo. Sejam $\boldsymbol{D} = (\boldsymbol{t}, \boldsymbol{z}, \boldsymbol{\delta})$ o conjunto de dados observado e $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^*, \boldsymbol{\beta})$ o vetor de parâmetros a ser estimado, a contribuição de cada indivíduo para a verossimilhança, $L_j(\boldsymbol{\lambda}|\boldsymbol{D})$, é dada pela função densidade se o indivíduo apresentou o evento de interesse e pela função de sobrevivência se o indivíduo foi censurado. Sendo assim temos

$$L_j(\boldsymbol{\lambda}|\boldsymbol{D}) \propto f_{\text{pop}}(t_j|\boldsymbol{\lambda}, \boldsymbol{D})^{\delta_j}\ S_{\text{pop}}(t_j|\boldsymbol{\lambda}, \boldsymbol{D})^{1-\delta_j},$$

o que nos permite concluir que a função de verossimilhança considerando todos os indivíduos observados é dada por

$$L(\boldsymbol{\lambda}|\boldsymbol{D}) \propto \prod_{j=1}^{n} f_{\text{pop}}(t_j|\boldsymbol{\lambda}, \boldsymbol{D})^{\delta_j}\ S_{\text{pop}}(t_j|\boldsymbol{\lambda}, \boldsymbol{D})^{1-\delta_j}, \tag{2.9}$$

e a função log-verossimilhança é dada por

$$l(\boldsymbol{\lambda}|\boldsymbol{D}) \propto \sum_{j=1}^{n} \log\left\{ f_{\text{pop}}(t_j|\boldsymbol{\lambda}, \boldsymbol{D})^{\delta_j}\ S_{\text{pop}}(t_j|\boldsymbol{\lambda}, \boldsymbol{D})^{1-\delta_j} \right\}. \tag{2.10}$$

2.2.2 Distribuições a priori e a posteriori

Assumimos que as distribuições *a priori* dos parâmetros ϕ, γ, ψ e $\boldsymbol{\beta}$ são próprias e independentes como assumido na Seção 1.2.2. Logo, a distribuição *a posteriori* de $\boldsymbol{\lambda} = (\phi, \gamma, \psi, \boldsymbol{\beta})$ é dada por

$$\pi(\boldsymbol{\lambda}|\boldsymbol{D}) \propto \exp\left\{l(\boldsymbol{\lambda}|\boldsymbol{D})\right\} \pi(\boldsymbol{\lambda}),$$

em que $l(\boldsymbol{\lambda}|\boldsymbol{D})$ e $\pi(\boldsymbol{\lambda})$ são dadas, respectivamente, por (2.10) e (1.12).

Para a implementação do algoritmo Metropolis-Hastings, descrito na Seção 1.2.2, precisamos das densidades condicionais completas de cada um dos parâmetros, que são apresentadas a seguir:

$$\pi(\phi|\gamma, \psi, \boldsymbol{\beta}, \boldsymbol{D}) \propto \phi^{a_\phi - 1} \mathrm{e}^{-b_\phi \phi} \left\{ \sum_{j=1}^{n} \sum_{i=1}^{m_j} [\log(1+\phi) - (1 + \phi t_{j_i} + \phi \frac{\gamma}{\gamma+1} x_{j_i}) x_{j_i}^{\gamma} \psi^{i-1}] \right\},$$

$$\pi(\gamma|\phi,\psi,\boldsymbol{\beta},\boldsymbol{D}) \propto \gamma^{a_\gamma-1}\exp\Bigg\{-b_\gamma\gamma+\sum_{j=1}^{n}\sum_{i=1}^{m_j}[\log\gamma+(\gamma-1)\log x_{i_j} - (1+\phi t_{j_i}+\phi\frac{\gamma}{\gamma+1}x_{j_i})x_{j_i}^{\gamma}\psi^{i-1}]\Bigg\}$$

$$\pi(\psi|\phi,\gamma,\boldsymbol{\beta},\boldsymbol{D}) \propto \psi^{a_\psi-1}\exp\Bigg\{-b_\psi\psi+\sum_{j=1}^{n}\sum_{i=1}^{m_j}[(i-1)\log\psi - (1+\phi t_{j_i}+\phi\frac{\gamma}{\gamma+1}x_{j_i})x_{j_i}^{\gamma}\psi^{i-1}]\Bigg\},$$

e

$$\pi(\beta_k|\phi,\gamma,\psi,\boldsymbol{\beta}_{-k},\boldsymbol{D}) \propto \exp\Bigg\{-\frac{1}{2\sigma_{\beta_k}^2}(\beta_k-\mu_{\beta_k})^2 + \sum_{j=1}^{n}\sum_{i=1}^{m_j}[\boldsymbol{\beta}^T\boldsymbol{z}_j-(1+\phi t_{j_i}+\phi\frac{\gamma}{\gamma+1}x_{j_i})x_{j_i}^{\gamma}\psi^{i-1}]\Bigg\},$$

em que a e b, indexados pelos parâmetros, são os parâmetros de forma e de escala da densidade gama das distribuições *a priori* de ϕ,γ e ψ; μ_{β_k} e σ_{β_k} são, respectivamente, as médias os desvios padrão das distribuições *a priori* de cada um dos β_k; e $\boldsymbol{\beta}_{-k} = (\beta_1,\ldots,\beta_{k-1},\beta_{k+1},\ldots,\beta_K)$, ou seja, é o vetor de parâmetros $\boldsymbol{\beta}$ sem a k-ésima componente.

Novamente, as densidades condicionais não nos remetem a nenhuma distribuição conhecida. Entretanto, o algoritmo Metropolis-Hastings é capaz de gerar amostras de ϕ,γ,ψ e β_k através das distribuições condicionais completas dos parâmetros desconhecidos, como mostra o esquema apresentado na Seção 1.2.2. Como a função de sobrevivência apresentada em (2.8) tem como casos particulares vários modelos, nosso interesse é verificar se um modelo mais simples poderia ser adotado. Devemos então testar hipóteses tais como $H_0:\theta=0$, $H_0:\psi=1$, $H_0:\phi=0$, $H_0:\gamma=1$, $H_0:\gamma=1,\psi=1$, $H_0:\phi=0,\psi=1$ e $H_0:\phi=0,\gamma=1$. Para isso utilizamos os critérios descritos na Seção 1.2.3.

2.2.3 Simulação

Os estudos de simulação foram realizados com o objetivo de analisar as probabilidades de cobertura (PC) dos intervalos de credibilidade e os procedimentos são análogos aos descritos na Seção 1.2.4. Os dados foram gerados baseados em um estudo clínico que visa analisar a eficácia de um tratamento aplicado a um grupo tratamento e um grupo controle. O vetor de parâmetros a ser estimado é dado por $\boldsymbol{\lambda}=(\phi,\gamma,\psi,\beta)$. Consideramos $\phi=2,2$, $\gamma=2,0$, $\psi=4,0$ e $\beta=-1,2$, diferentes tamanhos amostrais (30, 50, 70 e 100) e diferentes números de ocorrência (3 e 5). Para diferenciar os indivíduos pertencentes ao

Tabela 2.1: Probabilidades de cobertura dos intervalos de credibilidade de 95% para cada parâmetro e para cada tamanho amostral com $m = 3$ ocorrências para cada observação.

Parâmetro	Tamanho amostral			
	30	50	70	100
ϕ	0,817	0,848	0,856	0,920
γ	0,842	0,827	0,894	0,996
ψ	0,902	0,879	0,912	0,912
β	0,993	0,995	0,999	0,999

Tabela 2.2: Probabilidades de cobertura dos intervalos de credibilidade de 95% para cada parâmetro e para cada tamanho amostral com $m = 5$ ocorrências para cada observação.

Parâmetro	Tamanho amostral			
	30	50	70	100
ϕ	0,840	0,849	0,901	0,978
γ	0,915	0,891	0,919	0,989
ψ	0,843	0,905	0,953	0,957
β	0,995	0,992	0,990	0,992

grupo tratamento dos pertencentes ao grupo controle, utilizamos uma covariável binária z, tal que z é igual a -1 ou 1. A distribuição $\Gamma(0,9; 0,3)$, cuja média é 3 e a variância é 10, foi considerada como distribuição *a priori* dos parâmetros ϕ, γ e ψ. Para o parâmetro β foi considerada uma distribuição normal com média 0 e variância 100. Os resultados da PC para diferentes tamanhos amostrais e diferentes números de recorrência estão organizados nas Tabelas 2.1 e 2.2, que permitem concluir que números pequenos e moderados de recorrência assim como de observações no conjunto de dados não prejudicam as estimativas dos parâmetros.

2.3 Aplicação

Analisamos o conjunto de dados descrito na Seção 1.3. O grupo tratamento apresentou, em média, 3,5 tumores recorrentes, e o grupo controle apresentou um número médio maior, 5,7, o que evidencia a possível existência de uma fração de cura na população.

As distribuições *a priori* usadas foram $\phi \sim \Gamma(2,25; 1,5)$ o que nos dá $\mathrm{E}(\phi) = 1,5$ e

$\text{Var}(\phi) = 1$; $\gamma \sim \Gamma(0,9; 0,3)$, $\psi \sim \Gamma(0,9; 3)$, sendo $\text{E}(\gamma) = \text{E}(\psi) = 3$ e $\text{Var}(\gamma) = \text{Var}(\psi) = 10$; e $\beta \sim \mathcal{N}(0; 100)$. Os valores dos hiperparâmetros foram escolhidos subjetivamente, mas tendo em mente um balanço entre assegurar a não informação e a convergência do algoritmo Metropolis-Hastings. Foram geradas duas cadeias com 55000 iterações cada. As 5000 primeiras foram descartadas para reduzir a possível influência do ponto inicial da cadeia. As restantes foram selecionadas de 50 em 50 para evitar a correlação nas séries, resultando numa amostra com 2000 observações. Para a implementação do algoritmo, assim como a verificação da convergência das cadeias realizamos os mesmos procedimentos descritos na Seção 1.3.

A Tabela 2.3 apresenta as médias *a posteriori* e os respectivos desvios padrão (entre parênteses) dos parâmetros. A Tabela 2.4 apresenta os valores de três estatísticas AIC, BIC e DIC, descritos na Seção 1.2.3. Os resultados dão evidência positiva ao modelo completo, mostrando a importância de serem levadas em consideração a contagem do número de eventos e as duas escalas de tempo na análise. As Figuras 2.1 e 2.2 mostram os históricos das cadeias e as densidades *a posteriori* marginais para os parâmetros. A probabilidade de cura está associada com o parâmetro β por meio da função logística (2.7). Então a estimativa de β nos dá que o grupo tratamento tem aproximadamente 58% de probabilidade de cura, enquanto a probabilidade de cura do grupo controle é de 42%. Logo, há evidência de o tratamento ser benéfico. Tal conclusão está em concordância com as conclusões obtidas por Lawless (1995) e Louzada Neto (2008). No entanto, estes autores não estimaram a probabilidade de cura de cada grupo.

2.4 Comentários finais

Dados de sobrevivência com fração de cura são encontrados em diversas áreas, inclusive em situações em que a recorrência de eventos existe e não deve ser ignorada. O modelo proposto EMT com longa duração engloba ambas as escalas de tempo, a intervalar e o tempo total, a contagem de eventos, além de covariáveis e tem flexibilidade para incluir uma fração de cura. O modelo acomoda diversos casos particulares que podem ser testados diretamente. As estimativas dos parâmetros são obtidas utilizando um método completamente bayesiano, que permite a incorporação de conhecimentos prévios. Os resultados da simulação mostram a eficácia do método de estimação dos parâmetros tanto para uma

Tabela 2.3: Médias *a posteriori* e respectivos desvio padrão (entre parênteses).

Modelo	Parâmetro			
	γ	ϕ	ψ	β
EMT	1,911	1,786	1,105	0,253
	(0,234)	(0,961)	(0,073)	(0,102)
Modelo Weibull	1,283	-	1,303	0,137
ordinário com contagem	(0,175)	-	(0,083)	(0,246)
Processo de	0,981	1,170	-	0,132
Poisson-renovação	0,113	(0,713)	-	(0,247)
Processo de Poisson	-	0,411	1,153	0,117
não-homogêneo com contagem	-	(0,328)	(0,042)	(0,218)
Processo de contagem	-	-	1,172	0,686
	-	-	(0,047)	(0,253)
Modelo Weibull ordinário	0,831	-	-	0,776
	(0,086)	-	-	(0,257)
Processo de Poisson	-	1,257	-	0,150
não-homogêneo	-	(0,549)	-	(0,244)

quantidade pequena de indivíduos na amostra como para uma quantidade moderada na presença de censura.

Tabela 2.4: Valores dos critérios AIC, BIC e DIC

Modelo	AIC	BIC	DIC
EMT	-36,087	-36,228	-37,107
Modelo Weibull ordinário com contagem	-33,142	-33,401	-35,263
Processo de contagem	-28,125	-28,383	-31,743
Processo de Poisson não-homogêneo com contagem	-27,289	-27,547	-31,931
Processo de Poisson-renovação	-18,008	-18,265	-21,949
Processo de Poisson não-homogêneo	-17,942	-18,200	-22,873
Modelo Weibull ordinário	-14,452	-14,709	-18,388

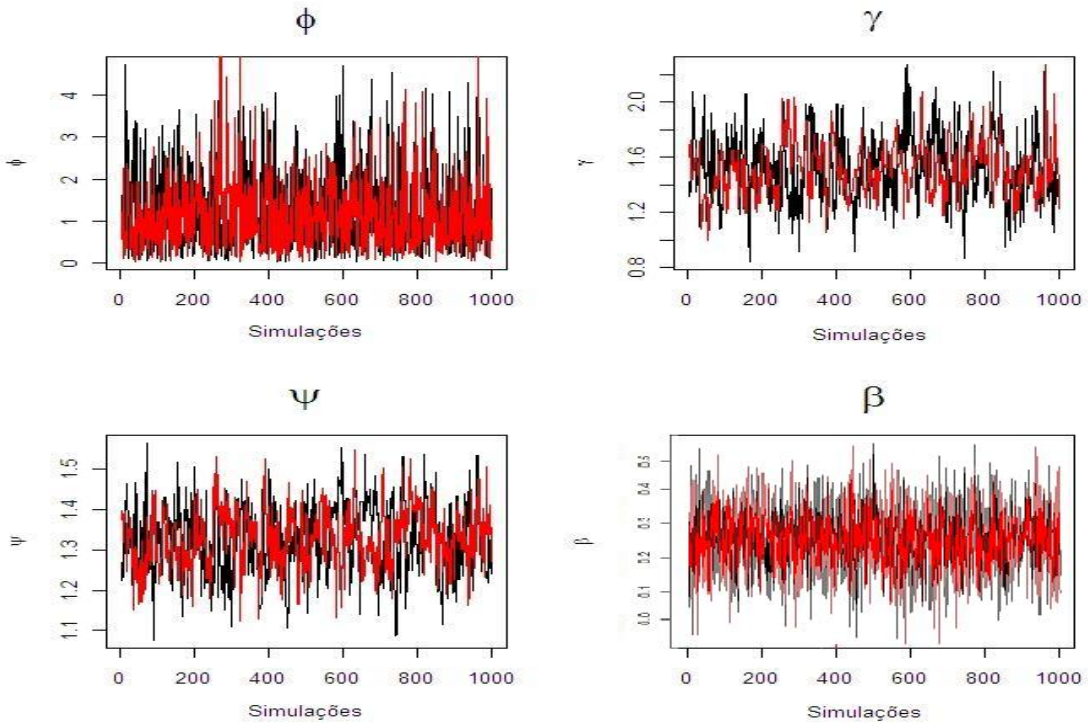

Figura 2.1: Histórico das cadeias.

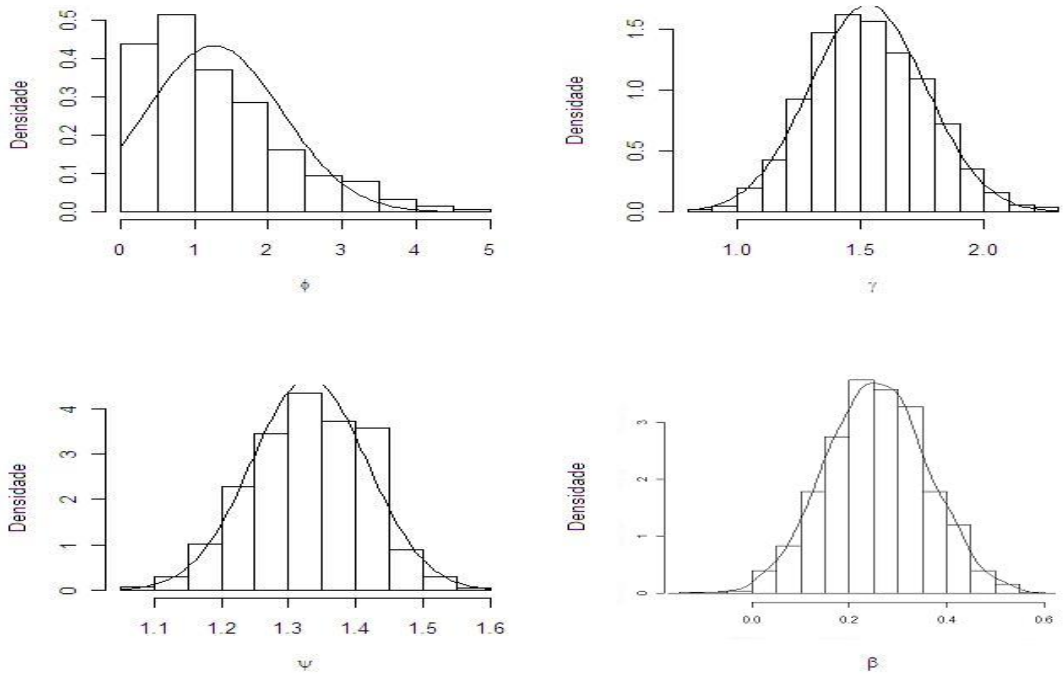

Figura 2.2: Densidades marginais *a posteriori*.

Capítulo 3

Modelo de Sobrevivência Geral na Presença de Causas Competitivas

Em muitos estudos de sobrevivência há uma parte da população que, devido a certa intervenção visando impedir a ocorrência do evento de interesse, pode vir a não ser suscetível a este evento, ou seja, uma porcentagem da população responde favoravelmente a tal intervenção, sendo considerada fora de risco. O modelo clássico de Berkson-Gage (Berkson & Gage, 1952), estudado por Farewell (1982, 1986), Goldman (1984), Sy & Taylor (2000), Banerjee & Carlin (2004), entre muitos outros, assim como modelos mais recentes e abrangentes (Yakovlev & Tsodikov, 1996; Chen, Ibrahim & Sinha, 1999; Ibrahim *et al.*, 2001; Chen, Ibrahim & Sinha, 2002; Yin & Ibrahim, 2005) incorporam a possibilidade de avaliar a população curada de diversas formas.

Muitas pesquisas ignoram a possível existência da proporção de indivíduos fora de risco censurando-os, já que em modelos de sobrevivência paramétricos tradicionais os indivíduos que não apresentaram o evento de interesse até determinado tempo são considerados censurados. Nos estudos de sobrevivência, a existência de uma proporção de indivíduos fora de risco chamada de longa duração ou de fração de cura. O último termo é mais comum em estudos da área médica ou biológica e será empregado neste trabalho, já que nossa motivação é dada na área médica, mais especificamente em estudos oncológicos.

A ocorrência do evento de interesse pode ser dada por uma ou várias causas competitivas (Gordon, 1990). O número de causas, assim como o tempo de sobrevivência associado a cada causa, não são observados (Cox & Oakes, 1984) e são chamados de fatores ou riscos latentes. O modelo proposto por Chen *et al.* (1999) baseia-se na existência

de fração de cura com fatores latentes, assim como, por exemplo, Yakovlev & Tsodikov (1996) Ibrahim *et al.* (2001), Chen *et al.* (2002), Banerjee & Carlin (2004) e Yin & Ibrahim (2005). Uma outra abordagem é desenvolvida por Cooner, Banerjee, Carlin & Sinha (2007) que modelam estocasticamente a sequência ordenada de tempos latentes, os quais induzem a ocorrência do evento em estudo. Para mais detalhes veja, por exemplo, Rodrigues, Cancho & de Castro (2008) e Balka, Desmond & McNicholas (2009). O cenário de causas competitivas permite longa duração quando a probabilidade de o número de causas latentes ser igual a zero é não nula.

O número de riscos latentes pode ser modelado por qualquer distribuição com média positiva e finita e suporte discreto, por exemplo, as distribuições de Poisson, binomial negativa, geométrica, de Bernoulli, COM-Poisson (Chen *et al.*, 1999; de Castro, Cancho & Rodrigues, 2007; Cooner *et al.*, 2007; Rodrigues, de Castro, Cancho & Balakrishnan, 2009b; Rodrigues, Cancho, de Castro & Louzada-Neto, 2009a). O modelo de Berkson-Gage (Berkson & Gage, 1952) pode ser considerado como um desses casos em que o número de riscos latentes tem distribuição de Bernoulli e há no máximo um risco latente.

O modelo proposto é motivado por estudos clínicos na área oncológica que objetivam, por exemplo, detectar a eficácia de um tratamento. Vale ressaltar que uma das funções do sistema linfático é transportar células imunes de um para outro linfonodo. Se os linfonodos falham, podem transportar células cancerígenas causando um processo chamado de metástase (Pollock, Doroshow, Khayat, Nakao & O'Sullivan, 2004). A gravidade da doença pode ser julgada pela quantidade de linfonodos contaminados, quanto mais linfonodos estiverem contaminados, maior é a preocupação com a metástase, e por isso mais agressiva é a cirurgia, ou seja, maior é a região a ser mutilada. No entanto, a retirada desses linfonodos pode diminuir a qualidade de vida da paciente. É por este motivo que objetivamos estudar a capacidade de contaminação dos linfonodos.

É comum as pesquisas nesta área abordarem o crescimento do tumor e não seu processo de metástase. Nesta situação, o tempo de desenvolvimento de um novo tumor é obtido pela presença de um número de células clonogênicas (clonogenes) que sobrevivem ao tratamento. Assim, cada clonogene sobrevivente é considerado um fator de risco. Diversos trabalhos supõem que o número de clonogenes sobreviventes segue uma distribuição de Poisson ou uma binomial. Esta suposição foi questionada por Tucker, Thames & Taylor (1990), já que a distribuição de Poisson não leva em consideração a proliferação natural

dos clonogenes, e nem considera que os clonogenes podem ser mortos pelo tratamento. Hanin (2001) considerou o número de clonogenes sobreviventes como o estágio final de um processo estocástico de nascimento e morte e obteve a distribuição binomial negativa generalizada (BNG). Uma das principais características da BNG é considerar a proliferação natural dos clonogenes e a eficácia de cada dose do tratamento. Hanin (2001) provou que esta distribuição converge para uma distribuição de Poisson se o número inicial de clonogenes for suficientemente grande e se a eficácia de cada dose do tratamento for praticamente nula. Ilustrações da distribuição de Hanin (2001) e aplicações da sua forma limite podem ser encontradas em Zaider, Zelefsky, Hanin, Tsodikov, Yakolev & Leibel (2001) e Hanin, Zaider & Yalovlev (2001).

Neste trabalho assumimos que os linfonodos agem como causas competitivas para espalhar o câncer pelo corpo. Nosso objetivo é estimar a taxa da proliferação de contaminação dos linfonodos e a probabilidade de o tratamento não destruir as células cancerígenas do linfonodo. Sendo assim, supomos que o número de riscos latentes segue uma distribuição BNG. A vantagem desta suposição é incorporar à análise o número de linfonodos contaminados inicialmente, a taxa de contaminação natural dos linfonodos e uma intervenção para a descontaminação dos linfonodos. Nossa proposta nos permite calcular a probabilidade de cura, além das taxas envolvidas no processo.

Podemos supor que o tempo de ocorrência de cada causa seja representado por diversas distribuições, por exemplo, a exponencial, a exponencial por partes, a Weibull e a log-logística, entre outras. Neste trabalho consideramos duas possibilidades para a distribuição do tempo de ocorrência, a distribuição Weibull e a log-logística.

Nas Seções 3.1 e 3.3 descreveremos respectivamente os modelos BNG e BNGW motivados por estudos clínicos na área oncológica. A Seção 3.2 construímos a função de verossimilhança do modelo BNG. O procedimento de estimação de máxima verossimilhança e um estudo de simulação estão descritos na Seção 3.4. A inferência bayesiana é descrita na Seção 3.5. A Seção 3.6 apresenta uma aplicação a um conjunto de dados artificiais. Na Seção 3.7 um conjunto de dados de câncer de mama ilustra a aplicabilidade do modelo proposto. Comentários finais estão na Seção 3.8.

3.1 Modelo binomial negativo generalizado

Seja M uma variável aleatória que nota o número de causas competitivas latentes relacionadas à contaminação de outro linfonodo. Dado M igual a m, sejam $X_l, l = 1, \ldots, m$, os tempos de ocorrência do evento de interesse devido à l-ésima causa competitiva. Consideramos que X_l são variáveis aleatórias independentes e identicamente distribuídas com função distribuição de parâmetro $\boldsymbol{\gamma}$, que independe de M, dada por $F(x|\boldsymbol{\gamma}) = 1 - S(x|\boldsymbol{\gamma})$, em que $S(x|\boldsymbol{\gamma})$ denota a função de sobrevivência. Assim, o tempo de sobrevivência para cada indivíduo é definido por

$$Y = \min\{X_1, \ldots, X_M\},$$

com $P(Y = \infty | M = 0) = 1$, ou seja, há uma fração da população não suscetível à ocorrência do evento, dada por p_0.

Sejam λ a taxa de proliferação natural dos linfonodos e η a probabilidade de cada dose do tratamento não destruir as células cancerígenas do linfonodo. O número de linfonodos contaminados inicialmente, i, assim como a quantidade de doses do tratamento, k, e o intervalo entre uma dose e outra igual, τ, são conhecidos. Consideramos que M segue uma distribuição BNG, cuja distribuição de probabilidade, $p_m = P(M = m), m = 0, 1, 2, \ldots$, é dada por (Hanin, 2001)

$$p_m = \left(\frac{a}{c}\right)^i \left(\frac{b}{a}\right)^m Q_m\left(\frac{ad - bc}{bc}\right), \quad m \geq 0, \tag{3.1}$$

em que $Q_0(x) = 1$,

$$Q_m(x) = \sum_{r=1}^{m} \begin{pmatrix} m-1 \\ m-r \end{pmatrix} \begin{pmatrix} i+r-1 \\ r \end{pmatrix} x^r, \quad m \geq 1,$$

em que a, b, c e d são relacionados com os parâmetros de interesse $\lambda > 0$ e $0 \leq \eta \leq 1$ por (Hanin, 2001)

$$a = 1 - \omega - \eta + \eta\omega\mu^{k-1},\ b = \eta(\omega\mu^{k-1} - 1),\ c = 1 - \omega - \eta + \eta\mu^{k-1} \text{ e } d = \eta(\mu^{k-1} - 1), \tag{3.2}$$

em que

$$\alpha = \mathrm{e}^{-\lambda\tau}, \quad \omega = \frac{\lambda - \eta\lambda}{\lambda(1 - \alpha)} \quad \text{e} \quad \mu = \frac{\eta}{\alpha} = \eta\ \mathrm{e}^{\lambda\tau}, \tag{3.3}$$

A função geradora de probabilidades é dada por

$$A(s) = \sum_{m=0}^{\infty} p_m s^m = \left(\frac{a - bs}{c - ds}\right)^i, \quad 0 \leq s \leq 1. \tag{3.4}$$

Sejam $S_{\text{pop}}(y) = P(Y > y)$ a função sobrevivência da variável aleatória Y e $S(x|\boldsymbol{\gamma})$ a função de sobrevivência associada à $X_l, l = 1, \ldots, M$, temos que

$$S_{\text{pop}}(y) = P(M = 0) + P(X_1 >, X_2 > y, \ldots, X_M > y, M \geq 1).$$

Rodrigues *et al.* (2009a), entre outros, provaram que

$$S_{\text{pop}}(y) = p_0 + \sum_{m=1}^{\infty} p_m S^m(y|\boldsymbol{\gamma}) = A(S(y|\boldsymbol{\gamma}))$$

e $S_{\text{pop}}(\infty) = p_0$, sendo $A(\cdot)$ a função geradora de probabilidades do número de causas e, por definição, $S_{\text{pop}}(\infty) = \lim_{y\to\infty} S_{\text{pop}}(y)$. Assim, a função de sobrevivência imprópria é dada por

$$S_{\text{pop}}(y) = \left(\frac{a - bS(y|\boldsymbol{\gamma})}{c - dS(y|\boldsymbol{\gamma})}\right)^i. \tag{3.5}$$

As correspondentes funções densidade e de risco impróprias são

$$f_{\text{pop}}(y) = -S'_{\text{pop}}(y) = A'(S(y|\boldsymbol{\gamma}))f(y|\boldsymbol{\gamma}) = \frac{if(y|\boldsymbol{\gamma})(ad - bc)}{(c - dS(y|\boldsymbol{\gamma}))^2}\left(\frac{a - bS(y|\boldsymbol{\gamma})}{c - dS(y|\boldsymbol{\gamma})}\right)^{i-1} \tag{3.6}$$

e

$$h_{\text{pop}}(y) = \frac{f_{\text{pop}}(y)}{S_{\text{pop}}(y)} = \frac{if(y)(ad - bc)}{(a - bS(y|\boldsymbol{\gamma}))(c - dS(y|\boldsymbol{\gamma}))},$$

em que $A'(s) = \sum_{m=1}^{\infty} mp_m s^{m-1}$. A fração de cura é dada por

$$p_0 = \lim_{y\to\infty} S_{\text{pop}}(y) = \lim_{y\to\infty}\left(\frac{a - bS(y|\boldsymbol{\gamma})}{c - dS(y|\boldsymbol{\gamma})}\right)^i = \left(\frac{a}{c}\right)^i,$$

já que $S(y|\boldsymbol{\gamma})$ é uma função de sobrevivência própria. Vale ressaltar que a fração de cura depende, além do número inicial de eventos ocorridos, do número de doses do tratamento, veja expressão (3.2).

3.2 Função de verossimilhança

É comum que dados de análise de sobrevivência não sejam completamente observados, ou seja, contenham censuras. Consideramos este caso com censura à direita. Seja C_j o tempo de censura do j-ésimo indivíduo. T_j denota o tempo de vida observado para o j-ésimo indivíduo e é dado por $T_j = \min\{Y_j, C_j\}$ e δ_j é a variável indicadora de censura tal que $\delta_j = 1$ se $Y_j \leq C_j$, e $\delta_j = 0$, caso contrário. Além disso, seja i_j o número inicial de linfonodos contaminados do j-ésimo indivíduo. Dessa forma, considerando que foram observados

n indivíduos, nosso conjunto de dados é formado pelos vetores $\boldsymbol{t} = (t_1, \ldots, t_n)^T, \boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^T$ e $\boldsymbol{i} = (i_1, \ldots, i_n)^T$. A correspondente função de verossimilhança é dada por

$$L(\lambda, \eta, \boldsymbol{\gamma}|\ \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{i}) \propto \prod_{j=1}^{n} f(t_j, \delta_j, i_j|\ \lambda, \eta, \boldsymbol{\gamma}) \tag{3.7}$$

em que

$$f(t_j, \delta_j, i_j|\ \lambda, \eta, \boldsymbol{\gamma}) = \sum_{m_j=0}^{\infty} S(t_j|\boldsymbol{\gamma})^{m_j - \delta_j} [m_j\ f(t_j|\boldsymbol{\gamma})]^{\delta_j}\ p_{m_j}. \tag{3.8}$$

A função de verossimilhança (3.7) pode ser reescrita como

$$L(\lambda, \eta, \boldsymbol{\gamma}|\boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{i}) \propto \prod_{j=1}^{n} f_{\text{pop}}(t_j)^{\delta_j}\ S_{\text{pop}}(t_j)^{1-\delta_j}. \tag{3.9}$$

De acordo com as expressões (3.5) e (3.6) temos finalmente que

$$L(\lambda, \eta, \boldsymbol{\gamma}|\ \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{i}) \propto \prod_{j=1}^{n} \left[i_j\ f(t_j|\boldsymbol{\gamma}) \frac{ad - bc}{(c - dS(t_j|\boldsymbol{\gamma}))^2} \right]^{\delta_j} \left[\frac{a - bS(t_j|\boldsymbol{\gamma})}{c - dS(t_j|\boldsymbol{\gamma})} \right]^{i_j - \delta_j}, \tag{3.10}$$

em que, a, b, c e d são dados por (3.2), $S(\cdot|\boldsymbol{\gamma})$ e $f(\cdot)$ são, respectivamente, a função de sobrevivência e a função densidade de probabilidade associada a cada causa X_l, $l = 0, \ldots, m$.

3.3 Modelo binomial negativo generalizado Weibull

Uma das possíveis distribuições para o tempo de ocorrência de cada causa é a distribuição Weibull, cuja distribuição e função densidade são dadas, respectivamente, por

$$F(x|\gamma_1, \gamma_2) = 1 - \exp(-x^{\gamma_1} \mathrm{e}^{\gamma_2}) \quad \text{e} \quad f(x|\gamma_1, \gamma_2) = \gamma_1 x^{\gamma_1 - 1} \exp(\gamma_2 - x^{\gamma_1} \mathrm{e}^{\gamma_2}), \tag{3.11}$$

em que $\gamma_1 > 0$ e $\gamma_2 \in \mathbb{R}$. A distribuição Weibull é uma das mais amplamente usadas para representar tempos de vida na análise de sobrevivência devido a sua versatilidade. Dependendo do valor de seu parâmetro de forma, γ_1, a distribuição Weibull é capaz de modelar uma variedade de comportamentos de vida. Sua taxa de falha é monótona decrescente para $\gamma_1 < 1$, para $\gamma_1 > 1$ é monótona crescente e para $\gamma_1 = 1$ é constante, equivalendo à distribuição exponencial, veja Figura 3.1 (Lawless, 2002).

Existem diversas formas de inclusão de covariáveis no modelo. Sugerimos que as covariáveis estejam diretamente relacionadas com um dos parâmetros da distribuição do tempo de ocorrência de cada causa dado, no nosso modelo, por γ_2. Desta forma, teremos

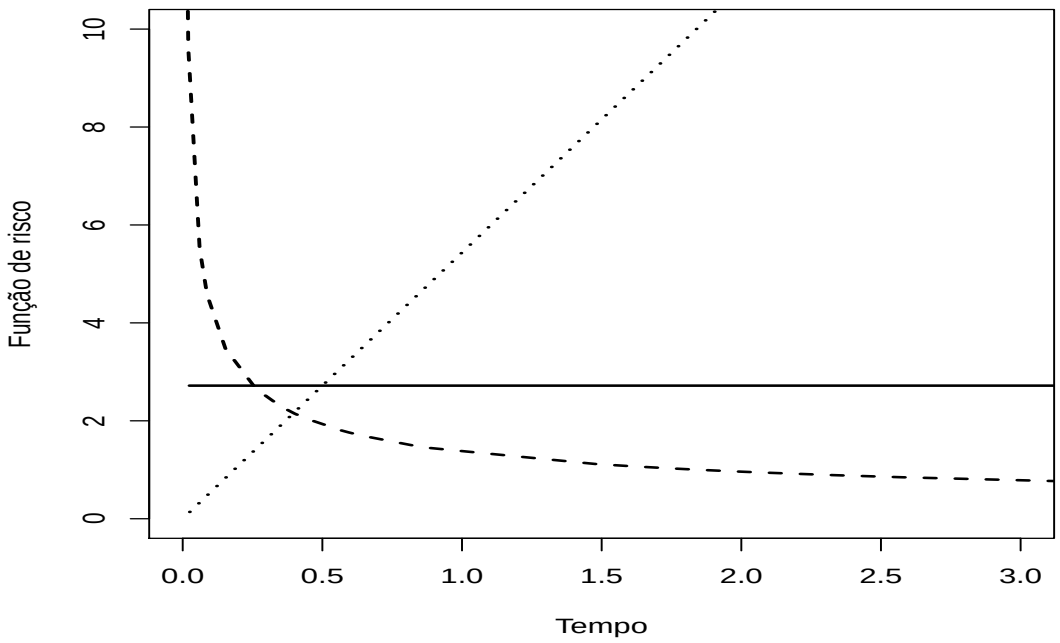

Figura 3.1: Função de risco da distribuição Weibull para $\gamma_1 = 0,5$ (- - -), $\gamma_1 = 1,0$ (—) e $\gamma_1 = 2,0(\cdots)$.

$\gamma_{2j} = \boldsymbol{\beta}^T \boldsymbol{z}_j$, $j = 1, \ldots, n$, em que $\boldsymbol{z}_j$ é o vetor de covariáveis para o j-ésimo indivíduo e $\boldsymbol{\beta}$ o vetor de parâmetros desconhecido sem intercepto. Nesta proposta, se $\beta_r > 0$ e o valor da covariável aumenta, então há aumento do risco. Já $\beta_r < 0$ e o aumento do valor da covariável z_r implicam na diminuição do risco.

Assim temos um modelo que acomoda causas competitivas latentes, em que o número de causas segue uma distribuição BNG e o tempo de cada causa segue uma distribuição Weibull. Tal modelo será chamado de modelo binomial negativo generalizado Weibull, ou simplesmente BNGW.

A função de verossimilhança do modelo BNGW é obtida combinando as expressões (3.10) e (3.11), o que nos dá

$$L(\lambda, \eta, \gamma_1, \boldsymbol{\beta} | \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{i}) = \prod_{j=1}^{n} \left\{ i_j \ \gamma_1 t_j^{\gamma_1 - 1} \exp(\gamma_{2j} - t_j^{\gamma_1} \mathrm{e}^{\gamma_{2j}}) \frac{ad - bc}{\left[c - d\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})\right]^2} \right\}^{\delta_j} \times \prod_{j=1}^{n} \left[\frac{a - b\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})}{c - d\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})} \right]^{i_j - \delta_j}, \quad (3.12)$$

em que $\gamma_{2j} = \boldsymbol{\beta}^T \boldsymbol{z}_j$, $j = 1, \ldots, n$.

3.4 Estimação de máxima verossimilhança

As estimativas de máxima verossimilhança (emv), $\widehat{\boldsymbol{\psi}} = (\widehat{\lambda}, \widehat{\eta}, \widehat{\boldsymbol{\gamma}})$, podem ser obtidas com a solução de $\boldsymbol{U}(\boldsymbol{\psi}) = \mathbf{0}$, em que

$$\boldsymbol{U}(\boldsymbol{\psi}) = \frac{\partial \log L(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \tag{3.13}$$

é o vetor escore. Sob certas condições de regularidades, $\widehat{\boldsymbol{\psi}}$ tem distribuição assintótica normal multivariada, $\mathcal{N}(\boldsymbol{\psi}, \boldsymbol{I}^{-1}(\boldsymbol{\psi}))$, em que

$$\boldsymbol{I}(\boldsymbol{\psi}) = \mathrm{E}\left(-\frac{\partial^2 \log L(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T}\right) \tag{3.14}$$

é a matriz informação de Fisher. Além disso $\boldsymbol{I}_{\mathrm{o}}(\boldsymbol{\psi}) = -\frac{\partial^2 \log L(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T}|_{\psi=\widehat{\psi}}$, chamada de matriz informação de Fisher observada, é um estimador consistente de $\boldsymbol{I}(\boldsymbol{\psi})$.

Apesar de a distribuição BNG ser mais flexível do que a distribuição de Poisson, devemos investigar qual delas fornece o melhor ajuste. Seja $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi_2})$ uma partição de $\boldsymbol{\psi}$. É possível mostrar que sob certas condições de regularidade e sob a hipótese $H_0 : \boldsymbol{\psi}_1 = \boldsymbol{\psi}_0$ a estatística da razão de verossimilhanças (RV)

$$\Lambda = -2 \log\left(\frac{L(\widehat{\boldsymbol{\psi}}_0)}{L(\widehat{\boldsymbol{\psi}})}\right) \tag{3.15}$$

tem distribuição assintoticamente dada por $\mathcal{X}^2_{(p)}$, em que p é a dimensão de $\boldsymbol{\psi}_1$ e $\widehat{\boldsymbol{\psi}}_0$ maximiza a verossimilhança sob H_0 (Lawless, 2002).

Nosso interesse principal é testar o modelo usual, dado pela Poisson (hipótese nula), contra um modelo mais abrangente dado pela BNG (hipótese alternativa). A distribuição de Poisson é um caso particular da BNG na fronteira do espaço paramétrico, com $\lambda \to 0$, o que faz com que as condições de regularidade não sejam satisfeitas e, então, a aproximação por uma distribuição qui-quadrado não necessariamente se aplica. Para avaliar a viabilidade desta aproximação, construímos diversas réplicas de conjuntos de dados sob o modelo Poisson e comparamos a estatística da RV com a distribuição $\mathcal{X}^2_{(1)}$. É esperado que a taxa de rejeição da hipótese nula não ultrapasse o nível de significância do teste. Além disso, analisamos o poder do teste da razão de verossimilhanças simulando conjuntos de dados construídos sob a hipótese alternativa para diferentes valores de λ, e nessa situação espera-se que a taxa de rejeição da hipótese nula tenda a um quando o número de réplicas tende a infinito (Davison & Hinkley, 1997). Os resultados serão apresentados na próxima seção.

Alternativamente, para comparar o modelo BNG com seus casos especiais, podemos considerar o AIC (Critério de informação Akaike) e o BIC (Critério de informação bayesiano), definidos, respectivamente, por $-2l(\widehat{\boldsymbol{\psi}_\rho}) + 2g$ e $-2l(\widehat{\boldsymbol{\psi}_\rho}) + g\log(n)$, em que $\widehat{\boldsymbol{\psi}_\rho}$ é a estimativa de máxima verossimilhança sobre o modelo ρ, g é o número de parâmetros estimados sobre o modelo g, e n é o tamanho amostral. Melhores modelos correspondem a menores valores de AIC e BIC.

3.4.1 Estudo de simulação

Com o intuito de verificar se o procedimento assintótico é válido para pequenos e moderados tamanhos de amostras, $n = 30$, 50, 70 e 100, foram efetuados diversos estudos de simulação. Assumimos que o tempo de ocorrência de cada causa segue uma distribuição Weibull com parâmetro $\gamma_1 = 0{,}5$; 1,0; 2,5, denotando os três comportamentos da taxa de falha da distribuição Weibull, como descrito na Seção 3.3. Além disso, assumimos que as covariáveis estão ligadas ao tempo de ocorrência de cada causa em função de duas covariáveis binárias, ou seja, $\gamma_{2j} = \boldsymbol{\beta}^T \boldsymbol{z}_j = \beta_1 z_{1_j} + \beta_2 z_{2_j}, j = 1, \ldots, n$, para diferentes tamanhos amostrais, $n = 30$, 50, 70 e 100. Para as simulações escolhemos o parâmetro ξ de forma a termos aproximadamente $(p_0 + 0,05)100\%$ de dados censurados, o que se enquadra no contexto de modelos com fração de cura. Consideramos $\lambda = 1,2$, $\eta = 0,6$ e $\boldsymbol{\beta} = (-1; -0,5)$, sob a hipótese de que o tratamento é aplicado em cinco doses, $k = 5$, espaçadas por uma unidade de tempo, $\tau = 1$. Nas simulações consideramos que os indivíduos apresentam inicialmente a mesma quantidade de linfonodos contaminados, ou seja, $i_j = i = 2$, $j = 1, \ldots, n$. A probabilidade de cura considerando estes valores para os parâmetros é aproximadamente igual a 30%. As simulações são descritas a seguir.

1. Gerar $u_j \sim \mathrm{U}(0, 1)$.
2. Se $u_j < p_0$, então $y_j = \infty$. Caso contrário, $y_j = F^{-1}\left(1 - \dfrac{a - cu_j^{1/i_j}}{b - du_j^{1/i_j}}\right)$, em que $F^{-1}(\cdot) = \left(-\dfrac{\log(\cdot)}{\exp(\gamma_{2j})}\right)^{1/\gamma_1}$.
3. Gerar $c_j \sim \mathrm{Exp}(\xi)$.
4. Fazer $t_j = \min\{y_j, c_j\}$.
5. Se $y_j < c_j$, então $\delta_j = 1$, caso contrário, $\delta_j = 0$, $j = 1, \ldots, n$.

Tabela 3.1: Probabilidades de cobertura empíricas para os intervalos de confiança dos parâmetros de interesse para $\gamma_1 = 0,5$, representando a distribuição Weibull com taxa de falha decrescente, amplitude média dos intervalos de confiança (entre parênteses) e $n =$ 30, 50, 70 e 100.

Parâmetro	Tamanho amostral			
	30	50	70	100
λ	0,807 (2,023)	0,872 (1,298)	0,877 (1,561)	0,901 (0,960)
η	0,844 (0,300)	0,931 (0,279)	0,924 (0,336)	0,910 (0,223)
γ_1	0,824 (1,789)	0,918 (1,399)	0,917 (1,339)	0,920 (1,021)
β_1	0,915 (1,612)	0,935 (1,367)	0,946 (1,113)	0,947 (0,976)
β_2	0,919 (0,744)	0,927 (0,385)	0,943 (0,356)	0,943 (0,259)

Foram então realizadas 1000 simulações cuja proporção de dados censurados está entre 35 e 45%, e em cada uma foram calculadas as estimativas de máxima verossimilhança através do algoritmo numérico *optim*, implementado no sistema R (R Development Core Team, 2009). Como valores iniciais dos parâmetros tomamos 0,5 para η, e 1 para λ, β_1 e β_2. Descartamos as simulações que não convergiram. Também foi obtido o intervalo de confiança de 95% para cada parâmetro baseado na teoria assintótica e verificado se o intervalo de confiança continha o verdadeiro valor do parâmetro, com o objetivo de obtermos a probabilidade de cobertura (PC) dos intervalos de confiança para cada parâmetro.

As Tabelas 3.1, 3.2 e 3.3 apresentam as probabilidades de cobertura dos intervalos nas diversas simulações. Concluímos que uma quantidade menor de indivíduos nos estudos não prejudica significativamente os resultados. Tal conclusão é bastante satisfatória já que em diversas aplicações não é possível obter um conjunto de dados com grande número de observações.

A performance da estatística da razão de verossimilhanças dada em (3.15) foi testada com nível de significância nominal de 5%, na comparação do modelo BNGW contra o modelo Poisson. Para isso foram calculados o poder do teste e o tamanho do teste para diferentes valores do parâmetro λ, diferentes tamanhos amostrais, $n = 30$, 50, 100 e 200 e diferentes comportamento da taxa de falha da distribuição Weibull, considerando $\gamma_1 = 0,5$, 1,0 e 2,5. Em todos os casos foram geradas 1000 amostras. Os resultados apresentados nas Tabelas 3.1, 3.5 e 3.6 nos permitem concluir, como esperado, que quanto menor o

Tabela 3.2: Probabilidades de cobertura empíricas para os intervalos de confiança dos parâmetros de interesse para $\gamma_1 = 1,0$, representando a distribuição exponencial, amplitude média dos intervalos de confiança (entre parênteses) e $n = 30$, 50, 70 e 100.

Parâmetro	Tamanho amostral			
	30	50	70	100
λ	0,818 (1,799)	0,879 (1,497)	0,905 (1,167)	0,922 (0,964)
η	0,878 (0,331)	0,931 (0,329)	0,931 (0,270)	0,939 (0,41)
γ_1	0,850 (1,651)	0,905 (1,324)	0,922 (1,301)	0,933 (1,07)
β_1	0,921 (1,656)	0,933 (1,306)	0,938 (1,100)	0,941 (0,918)
β_2	0,922 (1,433)	0,923 (0,749)	0,935 (0,647)	0,947 (0,561)

Tabela 3.3: Probabilidades de cobertura empíricas para os intervalos de confiança dos parâmetros de interesse para $\gamma_1 = 2,5$, representando a distribuição Weibull com taxa de falha crescente, amplitude média dos intervalos de confiança (entre parênteses) e $n = 30$, 50, 70 e 100.

Parâmetro	Tamanho amostral			
	30	50	70	100
λ	0,866 (1,823)	0,903 (1,569)	0,913 (1,172)	0,939 (1,043)
η	0,941 (0,414)	0,951 (0,357)	0,948 (0,328)	0,939 (0,257)
γ_1	0,871 (2,168)	0,915 (1,641)	0,930 (1,366)	0,936 (1,089)
β_1	0,930 (1,937)	0,936 (1,480)	0,943 (1,398)	0,947 (0,971)
β_2	0,925 (2,637)	0,938 (2,030)	0,951 (1,638)	0,949 (1,418)

valor de λ, maior dificuldade há em distinguir entre um modelo e outro. Tal dificuldade é suprida com o aumento do número de observações. A taxa de rejeição da hipótese nula fica em torno de 5% quando a hipótese alternativa converge para o modelo Poisson, o que é esperado teoricamente. Analogamente, 1000 réplicas do conjunto de dados sob o modelo Poisson foram construídas para analisar o desempenho da estatística Λ na comparação do modelo Poisson contra o modelo proposto. A taxa de rejeição da hipótese nula ficou abaixo do nível de significância de 5% para $n = 30$, 50 e 100 e atingiu o nível de significância esperado teoricamente para $n = 200$, como mostra a Tabela 3.7.

Tabela 3.4: Taxas de rejeição na comparação do modelo BNGW contra o modelo Poisson a um nível de significância nominal de 5% e para $\gamma_1 = 0,5$ representando a distribuição Weibull com taxa de falha decrescente.

Tamanho amostral	$\lambda = 2$	$\lambda = 1$	$\lambda = 0,5$	$\lambda = 0,1$
30	0,994	0,695	0,101	0,022
50	0,996	0,914	0,154	0,024
100	0,999	0,993	0,275	0,032
200	0,999	0,999	0,522	0,045

Tabela 3.5: Taxas de rejeição na comparação do modelo BNGW contra o modelo Poisson a um nível de significância nominal de 5% e para $\gamma_1 = 1,0$ representando a distribuição exponencial.

Tamanho amostral	$\lambda = 2$	$\lambda = 1$	$\lambda = 0,5$	$\lambda = 0,1$
30	0,991	0,705	0,288	0,114
50	0,997	0,909	0,311	0,023
100	0,999	0,821	0,396	0,042
200	0,999	0,999	0,517	0,051

3.5 Inferência bayesiana

Como alternativa à inferência clássica dada pela maximização da função de verossimilhança, sugerimos a inferência bayesiana. Nesta abordagem, combinamos a função de verossimilhança com informações *a priori* obtendo a distribuição *a posteriori*. As estimativas dos parâmetros são então dadas pelas médias da distribuição *a posteriori*.

Uma das formas de assegurarmos que a distribuição *a posteriori* seja própria é considerando distribuições *a priori* próprias (Ibrahim *et al.*, 2001). Embora não seja necessário, por simplicidade, assumiremos que os parâmetros são independentes *a priori*. Os parâmetros têm distribuições *a priori* de acordo com o espaço paramétrico de cada um deles, o que significa que $\lambda \sim$ Log-normal(a_0, a_1), $\eta \sim \mathcal{B}(b_0, b_1)$, $\gamma_1 \sim \Gamma(c_0, c_1)$ e que as G componentes de $\boldsymbol{\beta}$ são independentes *a priori* e cada β_g tem distribuição *a priori* normal, $\mathcal{N}(\mu_{\beta_g}, \sigma^2_{\beta_g})$, em que $a_0, a_1, b_0, b_1, c_0, c_1, \mu_{\beta_g}, \sigma^2_{\beta_g}$, $g = 1, \ldots, G$, são hiperparâmetros conhecidos. A função de verossimilhança (3.12) juntamente com tais suposições nos fornecem

Tabela 3.6: Taxas de rejeição na comparação do modelo BNGW contra o modelo Poisson a um nível de significância nominal de 5% e para $\gamma_1 = 2,5$ representando a distribuição Weibull com taxa de falha crescente.

Tamanho amostral	$\lambda = 2$	$\lambda = 1$	$\lambda = 0,5$	$\lambda = 0,1$
30	0,986	0,634	0,115	0,013
50	0,994	0,783	0,166	0,027
100	0,999	0,987	0,242	0,046
200	0,999	0,999	0,533	0,044

Tabela 3.7: Taxa de rejeição da hipótese nula na comparação do modelo Poisson contra o modelo BNGW. Para n = 30, 50, 100 e 200. Em cada célula, o resultado à esquerda corresponde a $\gamma_1 = 0,5$, o resultado ao centro corresponde a $\gamma_1 = 1$ e o resultado à direita corresponde a $\gamma_1 = 2,5$, representando, respectivamente a distribuição Weibull com taxa de falha decrescente, a distribuição exponencial e a distribuição Weibull com taxa de falha crescente.

$n = 30$	$n = 50$	$n = 100$	$n = 200$
0,073/0,022/0,001	0,056/0,038/0,002	0,035/0,034/0,021	0,047/0,051/0,062

que a distribuição *a posteriori* é dada por

$$\pi(\lambda,\eta,\gamma_1,\boldsymbol{\beta}|\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{i}) \propto \prod_{j=1}^{n}\left\{ i_j\ \gamma_1 t_j^{\gamma_1-1}\exp(\gamma_{2j}-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})\frac{ad-bc}{\left[c-d\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})\right]^2}\right\}^{\delta_j}$$
$$\times\prod_{j=1}^{n}\left[\frac{a-b\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})}{c-d\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})}\right]^{i_j-\delta_j}\pi(\lambda|a_0,a_1)\pi(\eta|b_0,b_1)\pi(\gamma_1|c_0,c_1)\prod_{g=1}^{G}\pi(\beta_g|\mu_g,\sigma_g^2).$$

Entretanto, independente das distribuições *a priori* escolhidas, a distribuição *a posteriori* para o modelo BNGW é analiticamente intratável. Uma alternativa é usarmos os métodos de Monte Carlo com cadeias de Markov (MCMC), como por exemplo o amostrador de Gibbs e o algoritmo de Metropolis-Hastings (veja p. ex. Chib & Greenberg, 1995). Para a implementação do algoritmo são necessárias as distribuições condicionais completas *a posteriori* de todos os parâmetros, dadas por

$$\pi(\lambda|\eta,\gamma_1,\boldsymbol{\beta},\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{i}) \propto \prod_{j=1}^{n}\frac{[ad-bc]^{\delta_j}[a-b\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})]^{i_j-\delta_j}}{[c-d\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})]^{i_j+\delta_j}}\pi(\lambda|a_0,a_1),$$

$$\pi(\eta|\lambda,\gamma_1,\boldsymbol{\beta},\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{i}) \propto \prod_{j=1}^{n}\frac{[ad-bc]^{\delta_j}[a-b\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})]^{i_j-\delta_j}}{[c-d\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})]^{i_j+\delta_j}}\pi(\eta|b_0,b_1),$$

$$\pi(\gamma_1|\lambda,\eta,\boldsymbol{\beta},\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{i}) \propto \prod_{j=1}^{n} \frac{[\gamma_1 t_j^{\gamma_1-1}\exp(\gamma_{2j}-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})]^{\delta_j}[a-b\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})]^{i_j-\delta_j}}{[c-d\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})]^{i_j+\delta_j}}\pi(\gamma_1|c_0,c_1)$$

e

$$\pi(\beta_g|\lambda,\eta,\gamma_1,\boldsymbol{\beta}_{-g},\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{i}) \propto \prod_{j=1}^{n} \frac{\exp[\delta_j(\gamma_{2j}-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})][a-b\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})]^{i_j-\delta_j}}{[c-d\exp(-t_j^{\gamma_1}\mathrm{e}^{\gamma_{2j}})]^{i_j+\delta_j}}\pi(\beta_g|\mu_g,\sigma_g^2),$$

em que $\boldsymbol{\beta}_{-g}=(\beta_1,\ldots,\beta_{g-1},\beta_{g+1},\ldots,\beta_G)$, ou seja, o vetor $\boldsymbol{\beta}$ sem a g-ésima componente. Todas estas distribuições condicionais não são avaliadas de forma fechada. Então faremos uso do algoritmo Metropolis-Hastings.

3.5.1 Estudos de simulação

O objetivo do estudo de simulação descrito nesta Seção é analisar as propriedades frequentistas do processo de estimação proposta na Seção 3.5. Então, assumimos que o número de causas competitivas segue uma distribuição BNG com parâmetros $\lambda = 2,0$ e $\eta = 0,4$ e que os tempos de cada causa seguem uma distribuição Weibull com parâmetro de forma ligado a duas covariáveis $\gamma_{2_j} = \boldsymbol{\beta}^T\boldsymbol{z}_j = \beta_1 z_{1_j} + \beta_2 z_{2j},\ j = 1,\ldots,n$, sendo $\boldsymbol{\beta} = (-1,2;2,3)^T$, e diferentes comportamentos representados por diferentes valores do parâmetro γ_1, ou seja, $\gamma_1 = 0,5$, 1,0 e 2,5. Obtivemos amostras de diferentes tamanhos, n =30, 50, 70 e 100, sob as hipóteses de que o tratamento é dado em 5 doses espaçadas por uma unidade de tempo e de que o número inicial de eventos é igual a quatro para todos os sujeitos da amostra, ou seja, $i_j = i = 4, j = 1,\ldots,n$. Sob tais considerações, a probabilidade de cura é de aproximadamente 25%.

Utilizamos distribuições *a priori* independentes e não informativas para os parâmetros tais que $\lambda \sim$ Log-normal$(0;10)$, $\eta \sim \mathcal{B}(1,1)$, $\gamma_1 \sim \Gamma(1;0,001)$, $\beta_1 \sim \mathcal{N}(0;1000)$ e $\beta_2 \sim \mathcal{N}(0;1000)$. Para cada parâmetro foram geradas duas cadeias paralelas de tamanho amostral 11000. Eliminamos as primeiras 1000 amostras e as restantes foram selecionadas de 5 em 5, resultando numa amostra de tamanho 4000. Realizamos este procedimento 1000 vezes para cada valor amostral analisado. Todo o estudo foi implementado no sistema R usando o pacote BRugs package (Thomas, O'Hara, Spiegelhalter, Best, Lunn & Rice, 2007). As Tabelas 3.8, 3.9 e 3.10 apresentam as PCs e as amplitudes médias dos intervalos de credibilidade 95% para cada parâmetro. Podemos concluir que a amplitude média dos intervalos de confiança diminui com o aumento do tamanho amostral e as PCs pouco variam nesta situação.

Tabela 3.8: Probabilidades de cobertura empíricas para os intervalos de credibilidade dos parâmetros de interesse e amplitude média dos intervalos de credibilidade (entre parênteses) para $\gamma_1 = 0,5$ representando a distribuição Weibull com taxa de falha decrescente e n = 30, 50, 70 e 100.

Parâmetro	Tamanho amostral			
	30	50	70	100
λ	0,982 (2,73)	0,984 (2,53)	0,985 (2,36)	0,991 (2,35)
η	0,976 (0,51)	0,981 (0,48)	0,988 (0,47)	0,989 (0,45)
γ_1	0,984 (0,58)	0,983 (0,47)	0,987 (0,44)	0,985 (0,40)
β_1	0,981 (2,25)	0,986 (1,93)	0,987 (1,61)	0,991 (1,83)
β_2	0,989 (4,10)	0,985 (3,45)	0,991 (3,19)	0,988 (3,17)

3.6 Dados artificiais

Analisamos dados fictícios baseados em um estudo oncológico. Em casos de câncer de mama, assim como em outros casos de câncer, o tratamento quimioterápico ocorre antes e depois da cirurgia de retirada do tumor. Neste caso, o tratamento é bem sucedido quando implica na não evolução do câncer e na descontaminação dos linfonodos, permitindo assim que apenas sejam retirados os tumores, preservando a mama da paciente.

Para a construção dos dados supomos que 70 mulheres apresentam quatro linfonodos contaminados no início do tratamento, $i_j = 4,\ j = 1, \ldots, 70$. Todas são submetidas a um tratamento quimioterápico, aplicado em cinco doses, $k = 5$, mensais, $\tau = 1$, com eficácia de 60%, ou seja, a chance de o câncer sobreviver a cada dose do tratamento é 60%, $\eta = 40\%$. O câncer evolui a uma taxa de $\lambda = 2,0$ por mês. Supomos que os tempos de ocorrência das causas competitivas seguem uma distribuição Weibull com parâmetros $\gamma_1 = 2,5$ e $\gamma_2 = \boldsymbol{\beta} z_j^T$. Os parâmetros $\beta_1 = -1,2$ e $\beta_2 = 2,3$ estão associados, respectivamente, às covariáveis, divididas em dois níveis acima de 45 anos e abaixo de 46 anos, e pela ausência ou presença de neutropenia (disfunção sanguínea caracterizada por uma contagem anormal de neutrófilos). Para os valores escolhidos a probabilidade de cura é aproximadamente igual a $p_0 = 23,1\%$.

Os dados e as estimativas de máxima verossimilhança foram obtidos como feito no estudo de simulação. A Tabela 3.11 nos apresenta os resultados e nos permite concluir

Tabela 3.9: Probabilidades de cobertura empíricas para os intervalos de credibilidade dos parâmetros de interesse e amplitude média dos intervalos de credibilidade (entre parênteses) para $\gamma_1 = 1,0$ representando a distribuição exponencial e $n = 30$, 50, 70 e 100.

Parâmetro	Tamanho amostral			
	30	50	70	100
λ	0,981 (2,22)	0,984 (2,12)	0,990 (1,84)	0,987 (1,85)
η	0,992 (0,40)	0,989 (0,42)	0,987 (0,35)	0,988 (0,32)
γ_1	0,978 (1,00)	0,983 (0,88)	0,988 (0,73)	0,991 (0,71)
β_1	0,988 (2,02)	0,992 (1,75)	0,989 (1,35)	0,985 (1,38)
β_2	0,985 (3,42)	0,987 (3,02)	0,991 (2,84)	0,989 (2,64)

que a probabilidade de cura estimada, $\widehat{p_0}$, é de aproximadamente $24,5\%$. O desvio padrão de cada estimativa foi calculado através da matriz hessiana estimada. É válido observar na Tabela 3.11 que os valores estimados se assemelham aos verdadeiros valores dos parâmetros, assim como os intervalos de confiança de 95% contêm os verdadeiros valores dos parâmetros.

A inferência bayesiana seguiu as mesmas considerações da Seção 3.5.1. A Tabela 3.12 apresenta as médias *a posteriori*, o desvio padrão e os intervalos de credibilidade 95% para cada parâmetro do modelo. A probabilidade de cura estimada é igual a 19,6%. A Figura 3.2 ilustra os gráficos do ajuste de Kaplan-Meier e da sobrevivência estimada pelo modelo selecionado nos casos clássico e bayesiano.

Uma análise dos resultados apresentados nas Tabelas 3.11 e 3.12 nos permite concluir que as estimativas fornecidas pelas duas inferências não têm diferenças significativas. O mesmo podemos concluir sobre ambas as probabilidades de cura estimadas. Há apenas uma pequena diferença entre os desvios padrão obtidos por cada procedimento, acarretando na diferença entre as amplitudes dos intervalos de confiança e as dos intervalos de credibilidade.

3.7 Dados de câncer de mama

Segundo dados do Instituto Nacional de Câncer (INCA) (Brasil, 2010), o câncer de mama é o tipo de câncer com maior taxa de mortalidade do mundo, por isso é considerado um

Tabela 3.10: Probabilidades de cobertura empíricas para os intervalos de credibilidade dos parâmetros de interesse e amplitude média dos intervalos de credibilidade (entre parênteses) para $\gamma_1 = 2,5$ representando a distribuição Weibull com taxa de falha crescente e n = 30, 50, 70 e 100.

Parâmetro	Tamanho amostral			
	30	50	70	100
λ	0,974 (2,56)	0,976 (2,09)	0,981 (1,87)	0,982 (1,81)
η	0,968 (0,41)	0,968 (0,35)	0,983 (0,32)	0,974 (0,31)
γ_1	0,976 (2,91)	0,977 (2,22)	0,981 (1,90)	0,990 (1,83)
β_1	0,954 (2,38)	0,976 (1,74)	0,976 (1,46)	0,981 (1,37)
β_2	0,983 (3,82)	0,982 (3,08)	0,976 (2,70)	0,987 (2,59)

Tabela 3.11: Verdadeiros valores, estimativas de máxima verossimilhança, desvios padrão, e intervalos de confiança.

Parâmetro	Verdadeiro valor	Estimativa	Desvio padrão	IC 95%
λ	2,0	1,955	0,446	(1,082;2,829)
η	0,4	0,394	0,081	(0,237;0,553)
γ_1	2,5	2,569	0,418	(1,751;3,388)
β_1	-1,2	-1,259	0,379	(-2,002;-0,516)
β_2	2,3	2,070	0,523	(1,045; 3,095)

problema de saúde pública. Para 2010 são previstos 49000 novos casos de câncer de mama no Brasil (Brasil, 2010). Antes de a física médica tomar um desenvolvimento satisfatório, a chance de cura da paciente era praticamente nula, e a análise de sobrevivência usual era adequada para o estudo de sobrevivência da paciente. Com seu avanço, modelos com fração de cura tornaram-se necessários para a análise, como por exemplo o modelo apresentado neste trabalho.

Os linfonodos agem como causas competitivas para alastrar o câncer pelo corpo. Neste contexto o modelo BNGW permite analisar o efeito dos linfonodos como causas competitivas para predizer a sobrevivência da paciente. A vantagem do nosso modelo é que podemos descrever a taxa de proliferação dos linfonodos contaminados, a probabilidade de o linfonodo contaminado sobreviver a cada dose do tratamento (eficácia da dose), além de estimar quais são as covariáveis com efeito significativo.

Tabela 3.12: Verdadeiros valores, médias *a posteriori*, desvios padrão e intervalos de credibilidade.

Parâmetro	Verdadeiro valor	Estimativa	Desvio padrão	ICred 95%
λ	2,0	1,863	0,538	(0,583;2,765)
η	0,4	0,423	0,103	(0,274;0,685)
γ_1	2,5	2,528	0,492	(1,470;3,443)
β_1	-1,2	-1,217	0,437	(-2,044;-0,278)
β_2	2,3	1,955	0,695	(0,175; 3,161)

Analisamos os dados de 40 mulheres com câncer de mama tratadas no Hospital Universitário da Faculdade de Medicina de Ribeirão Preto, com carcinoma ductal invasivo, submetidas a tratamento quimioterápico de 2003 a 2006 e acompanhadas até 2008 (Gozzo, 2008). Além de outras informações a respeito de cada paciente, o conjunto de dados descreve: idade da paciente (em anos, média 50,6 e desvio padrão 6,0), quantidade de leucócitos (gb1, $\times 10000/\text{mm}^3$, média 7,5 e desvio padrão 2,1), superfície corporal (sup, em mm^2, média 1,6 e desvio padrão 0,1) e duas drogas utilizadas no tratamento *taxotere* (txt, em mg, média 121,9 e desvio padrão 8,1) e *epirubician* (epi, em mg, média 88,8 e desvio padrão 9,7). O tratamento quimioterápico, para estas pacientes, foi dado em 4 a 6 doses mensais. Temos 75% de dados censurados. Depois de serem tratadas, as pacientes passaram por cirurgia para a retirada da região atingida pelo câncer. No nosso estudo analisamos a sobrevida da paciente após a cirurgia.

Antes de ajustarmos o modelo devemos identificar o comportamento da função de risco dos tempos observados. Para isso utilizamos um método gráfico baseado no teste do tempo total (TTT) (Aarset, 1985). Na sua versão empírica o gráfico TTT é dado por $G(r/n) = [(\sum_{j=1}^{r} Y_{j:n}) - (n-r)Y_{r:n}]/(\sum_{j=1}^{r} Y_{j:n})$, em que $r = 1, \ldots, n$ e $Y_{j:n}$ representam as estatísticas de ordem da amostra. É provado que a função de risco cresce (decresce) se o gráfico TTT é côncavo (convexo), quando se aproxima de uma linha diagonal é constante e, se primeiramente sua curvatura é côncava e depois convexa, seu risco cresce e depois decresce. Embora o gráfico TTT seja apenas uma condição suficiente e não necessária para indicar o formato da função de risco, será utilizado como um indicador de seu comportamento. A Figura 3.3 apresenta o gráfico TTT dos dados de câncer de mama que indica uma função de risco crescente, podendo então ser representada pela

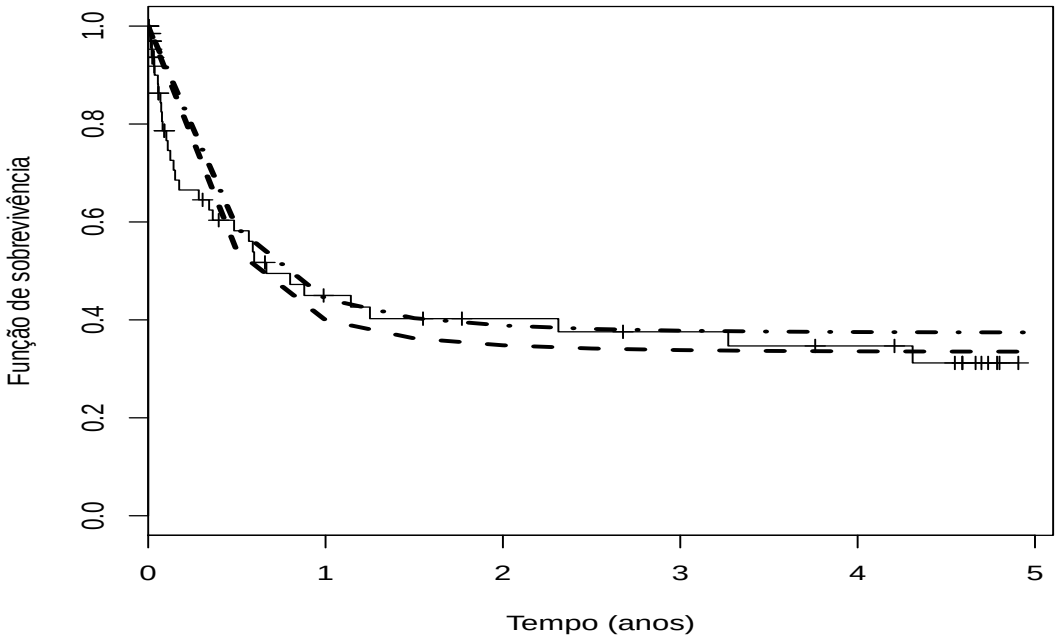

Figura 3.2: Estimativa de Kaplan-Meier e função de sobrevivência estimada pelo modelo BNGW via inferência clássica (- · - · -) e via inferência bayesiana (- - -) - dados artificiais.

distribuição Weibull.

Ajustamos nosso modelo BNGW e duas outras possibilidades: modelo de Poisson Weibull (PW) e modelo binomial negativo Weibull (BNW) ao conjunto de dados incluindo as cinco covariáveis descritas anteriormente. A Tabela 3.13 apresenta os valores de máximo da log-verossimilhança, $\max \log L(\cdot)$, e os valores das estatísticas AIC e BIC para os três modelos ajustados: PW, BNW e BNGW. As estatísticas AIC e BIC dão evidências a favor ao modelo BNGW. Os resultados das estimativas de máxima verossimilhança dos parâmetros do modelo BNGW, seus desvios padrão e seus intervalos de confiança 95% são apresentados na Tabela 3.14 e mostram que apenas as covariáveis *taxotere* e *epirubician* são significativas. As Tabelas 3.15 e 3.16 apresentam os resultados do ajuste que considera apenas as covariáveis significativas. Finalmente obtemos a probabilidade de cura estimada, $\widehat{p_0}$, para diferentes números de linfonodos contaminados inicialmente e para 4 ou 6 doses de tratamento (Tabela 3.17).

Também obtemos os ajustes para os modelo PW, BNW e BNGW através da inferência bayesiana. Utilizamos distribuições *a priori* independentes e não-informativas, sendo $\lambda \sim$ Log-normal$(0; 10)$, $\eta \sim \mathcal{B}(1, 1)$, $\gamma_1 \sim \Gamma(1; 0, 1)$ e para os βs distribuição normal $\mathcal{N}(0; 100)$.

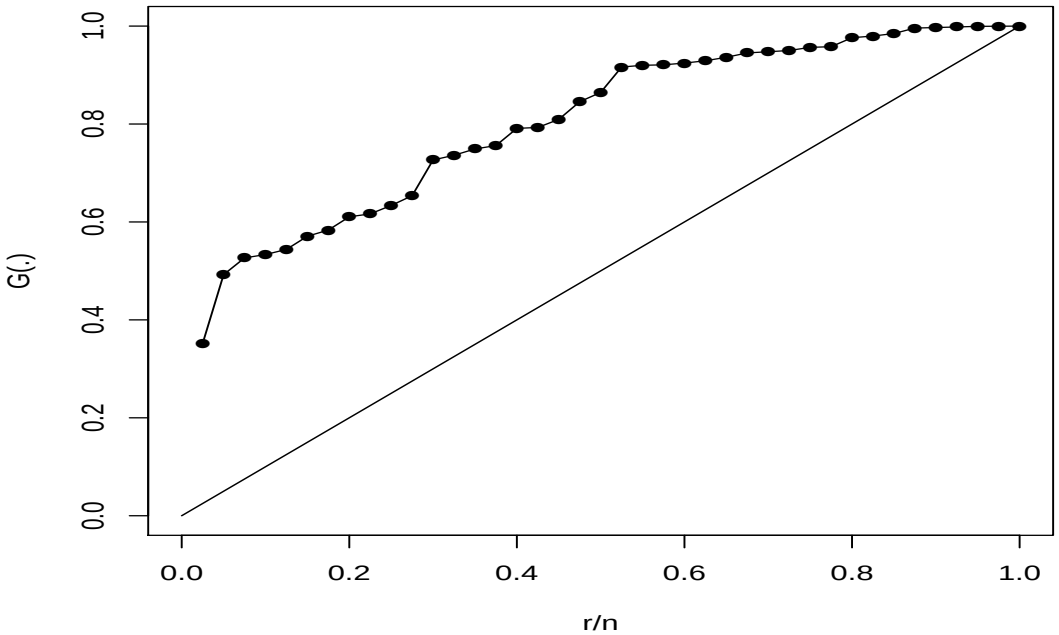

Figura 3.3: Gráfico TTTplot dos dados de câncer de mama.

Tabela 3.13: Valores $\max \log L(\cdot)$ e estatísticas AIC e BIC para os três modelos ajustados, PW, BNW e BNGW, considerando as cinco covariáveis.

Critério	PW	BNW	BNGW
$\max \log L(\cdot)$	-52,57	-52,60	-47,16
AIC	119,13	119,20	110,31
BIC	130,95	131,02	123,83

Geramos duas cadeias paralelas de tamanho 60000 para cada parâmetro. Descartamos as primeiras 10000 e as restantes selecionadas de 20 em 20, resultando numa amostra de tamanho 5000. Em uma das cadeias tomamos as estimativas de máxima verossimilhança como ponto inicial e na outra tomamos 0,5 para λ e para η, 1 para γ_1 e para 0 os βs. O código computacional foi implementado no OpenBUGS versão 3 (Spiegelhalter, Thomas, Best & Gilks, 1999) e os gráficos foram construídos com o auxílio do sistema R (R Development Core Team, 2009). A convergência das cadeias foi monitorada pela análise gráfica de Geweke (Geweke, 1992) e pelo método Gelman-Rubin (Brooks & Gelman, 1998) calculado pelo OpenBUGS.

As mesmas covariáveis abordadas na análise anterior foram incorporadas neste ajuste:

Tabela 3.14: Estimativas de máxima verossimilhança dos parâmetros do modelo BNGW, seus desvio padrão e seus intervalos de confiança assintóticos de 95% (IC 95%), considerando as cinco covariáveis.

Parâmetro	Estimativa	DP	IC 95%
λ	0,655	1,266	(0,123; 3,326)
η	0,615	1,077	(0,492; 0,725)
γ_1	3,605	0,302	(0,426; 30,488)
β_{txt}	-0,430	0,208	(-0,839; -0,022)
β_{gb1}	0,533	0,413	(-0,276; 1,343)
β_{idade}	-0,074	0,112	(-0,294; 0,146)
β_{epi}	0,396	0,177	(0,049; 0,743)
β_{sup}	1,116	13,796	(-25,924; 28,156)

Tabela 3.15: Valores $\max \log L(\cdot)$ e estatísticas AIC e BIC para os três modelos ajustados, PW, BNW e BNGW, considerando apenas as covariáveis significativas.

Critério	PW	BNW	BNGW
$\max \log L(\cdot)$	-92,32	-92,33	-90,37
AIC	192,74	192,47	193,09
BIC	199,48	199,22	202,08

taxotere, idade, *epirubician*, leucócitos e superfície corporal. Obtemos as estatísticas DIC para os modelos ajustados PW, BNW e BNGW, sendo seus respectivos valores iguais a 114,1, 119,4 e 108,4, o que favorece o modelo BNGW. A Tabela 3.18 apresenta as médias *a posteriori*, os desvios padrão e os intervalos de credibilidade para cada parâmetro. Os resultados indicam que apenas as covariáveis *taxotere* e *epirubician* são significativas e estão em concordância com a conclusão da análise anterior.

Em outro ajuste somente as covariáveis significativas foram incorporadas ao modelo. Então, os valores da estatística DIC dos modelos PW, BNW e BNGW são respectivamente iguais a 113,7, 118,4 e 107,9 dando evidências a favor do modelo BNGW. A Tabela 3.19 apresenta as médias *a posteriori* das amostras selecionadas para cada parâmetro, assim como os desvios padrão e os intervalos de credibilidade 95% das estimativas. As probabilidades de cura estimadas para diferentes números iniciais de linfonodos contaminados e diferentes quantidade de doses do tratamento estão descritas na Tabela 3.20. A

Tabela 3.16: Estimativas de máxima verossimilhança dos parâmetros do modelo BNGW, seus desvio padrão e seus intervalos de confiança assintóticos de 95% (IC 95%), considerando apenas as covariáveis significativas.

Parâmetro	Estimativa	DP	IC 95%
λ	0,604	0,894	(0,105; 3,482)
η	0,689	0,284	(0,559; 0,794)
γ_1	3,220	0,874	(0,580; 17,866)
β_{txt}	-0,298	0,103	(-0,501; -0,096)
β_{epi}	0,248	0,117	(0,019; 0,477)

Tabela 3.17: Probabilidade de cura, p_0, de acordo com o número inicial de linfonodos contaminados, i, e o número de doses, k, diferentes para cada paciente.

k	i					
	1	2	3	5	7	10
4	0,57	0,32	0,18	0,06	0,02	0,01
6	0,62	0,39	0,24	0,09	0,04	0,01

Figura 3.5 mostra o gráfico das amostras selecionadas para cada parâmetro e a Figura 3.6 apresenta as densidades marginais *a posteriori* aproximadas para cada parâmetro. Os gráficos do ajuste de Kaplan-Meier e da sobrevivência estimada pelo modelo selecionado no caso bayesiano são apresentados na Figura 3.7.

Baseados nos resultados apresentados obtidos extraímos diversas conclusões. Particularmente, λ denota a taxa de proliferação da contaminação, cuja estimativa é aproximadamente igual a 0,6 linfonodos por mês. O parâmetro η denota a probabilidade de a contaminação do linfonodo permanecer a cada dose do tratamento, com estimativa de aproximadamente 70%. Logo, assumindo doses independentes, a probabilidade de a contaminação no linfonodo resistir ao tratamento depois de 4 doses é igual a η^4, ou seja, $24,1\%$, e depois de 6 doses é igual a η^6, ou seja, $11,8\%$. Também podemos calcular a eficiência de cada dose definida por $1-\eta$, ou seja, cada dose independente tem 30% de eficácia. A probabilidade de cura depende do número de doses do tratamento, que se completo é dado em 6 doses. Então, receber o tratamento completo impacta positivamente na sobrevida da paciente, aumentando sua probabilidade de cura. A estimativa do parâmetro γ_1 é aproximadamente igual a 3, significando uma taxa de falha crescente

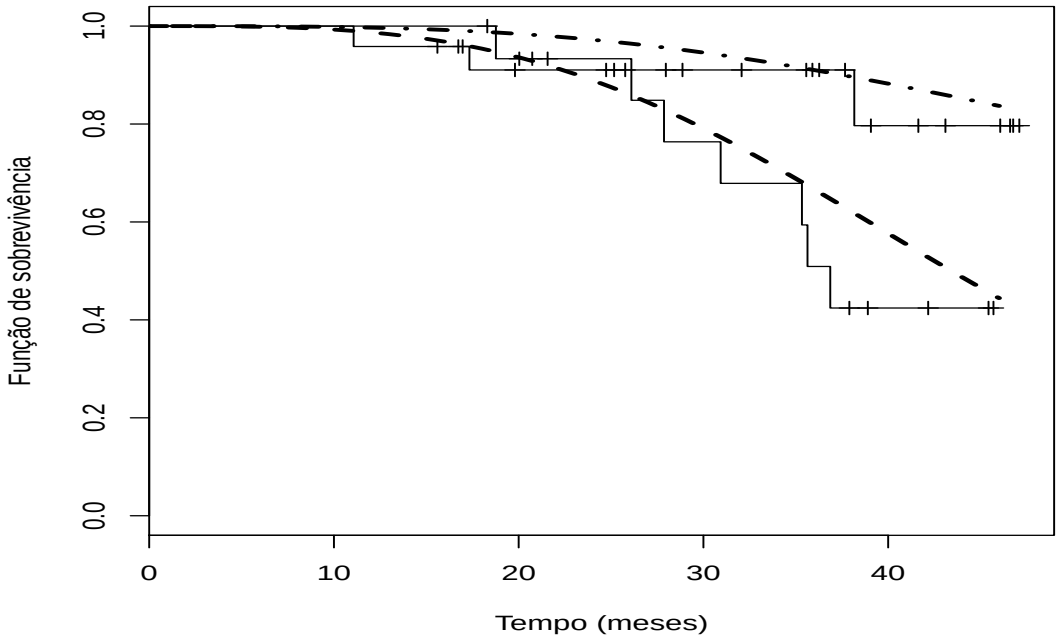

Figura 3.4: Estimativa de Kaplan-Meier e função de sobrevivência estimada pelo modelo BNGW via a inferência clássica: $i = 1$ (- · - · -) e $i > 1$ (- - -).

para uma distribuição Weibull, implicando que o risco de morte aumenta com o tempo, o que está em concordância com o gráfico TTT. Concluímos também que a droga *taxotere* produz aumento da sobrevivência o que está em acordo com Sparano, Wang, Martino, Jones, Perez, Saphner, Wolff & Sledge-Jr (2008).

3.8 Comentários finais

Neste capítulo propomos um modelo de sobrevivência geral para acomodar dados de sobrevivência na presença de causas competitivas latentes. Assumimos uma distribuição binomial negativa generalizada para o número de causas competitivas e uma distribuição Weibull para os tempos de ocorrência, obtendo o modelo BNGW. A principal vantagem de tal suposição é a estimação de duas importantes estatísticas num tratamento que visa combater o processo de metástase: a taxa de espalhamento da doença e a eficácia de cada dose do tratamento. Além disso, o modelo incorpora o número de doses, o intervalo de tempo entre as doses e a eficácia de cada dose. A relevância prática e a aplicabilidade do modelo foram demonstradas em um conjunto de dados reais de pacientes com câncer de

Tabela 3.18: Média *a posteriori* dos parâmetros do modelo BNGW, desvios padrão e intervalos de credibilidade 95% (ICred 95%), considerando as cinco covariáveis.

Parâmetro	Média *a posteriori*	DP	ICred 95%
λ	0,734	0,727	(0,003; 2,610)
η	0,623	0,173	(0,265; 0,917)
γ_1	3,624	1,036	(1,802; 5,974)
β_{txt}	-0,486	0,215	(-0,946; -0,108)
β_{gb1}	0,632	0,413	(-0,049; 1,580)
β_{idade}	-0,060	0,111	(-0,284; 0,156)
β_{epi}	0,494	0,231	(0,137; 1,026)
β_{sup}	-1,502	7,841	(-17,580; 13,08)

Tabela 3.19: Média *a posteriori* dos parâmetros do modelo BNGW, desvios padrão e intervalos de credibilidade 95% (ICred 95%), considerando apenas as covariáveis significativas.

Parâmetro	Média *a posteriori*	DP	ICred 95%
λ	0,540	0,569	(0,002; 1,980)
η	0,696	0,161	(0,334; 0,964)
γ_1	3,415	0,885	(1,911; 5,403)
β_{txt}	-0,333	0,115	(-0,601; -0,148)
β_{epi}	0,286	0,137	(0,075; 0,609)

mama. Do ponto de vista prático, o modelo proposto propiciou melhor ajuste aos dados, já que permitiu melhor interpretação e adaptação à situação.

Os dois processos de estimação apresentaram resultados similares, apesar da pequena quantidade de observações na amostra juntamente com a alta proporção de censura.

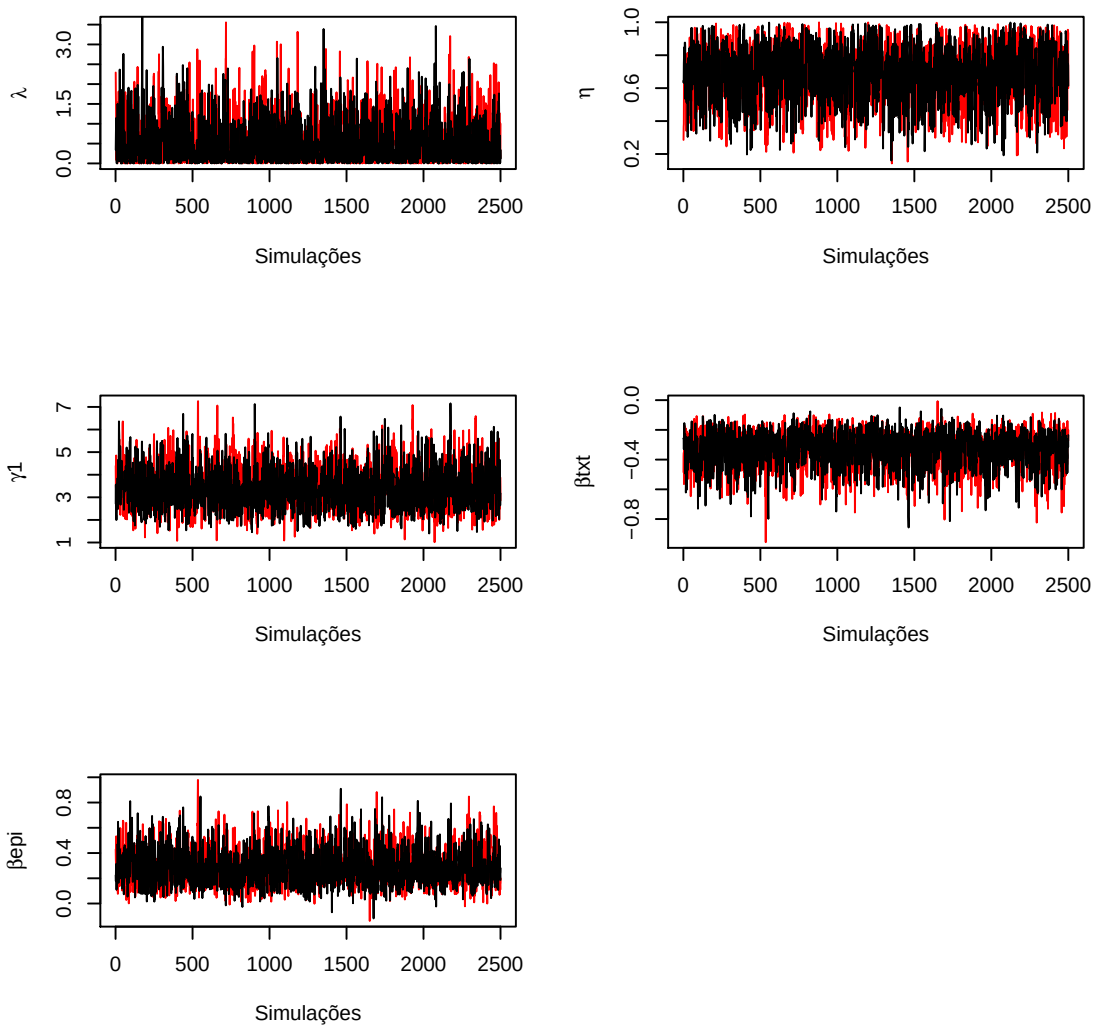

Figura 3.5: Histórico das cadeias.

Tabela 3.20: Probabilidade de cura, p_0, de acordo com o número inicial de linfonodos contaminados, i, e o número de doses, k, diferentes para cada paciente obtida pela inferência bayesiana.

k	i					
	1	2	3	5	7	10
4	0,58	0,33	0,19	0,06	0,02	0,01
6	0,64	0,41	0,26	0,11	0,04	0,01

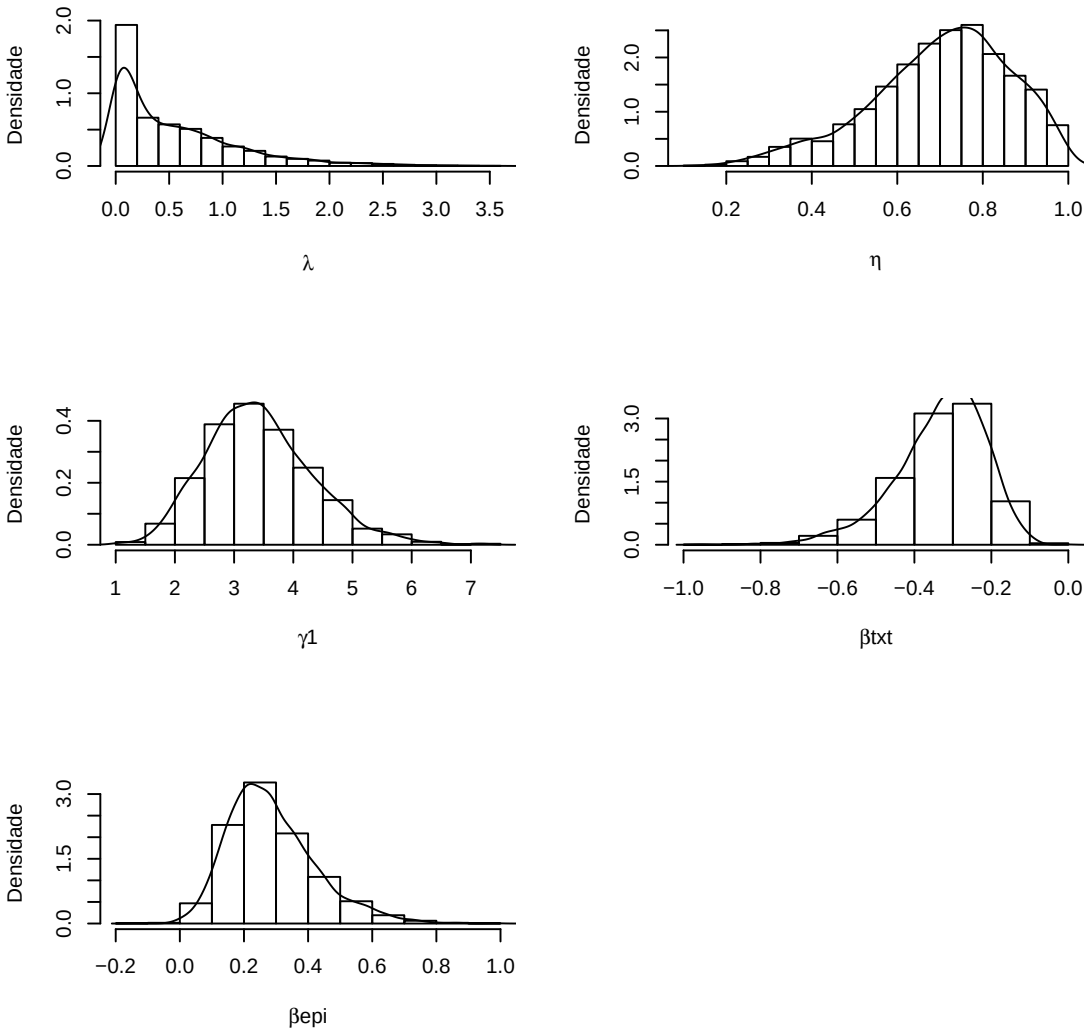

Figura 3.6: Densidades *a posteriori* marginais aproximadas dos parâmetros.

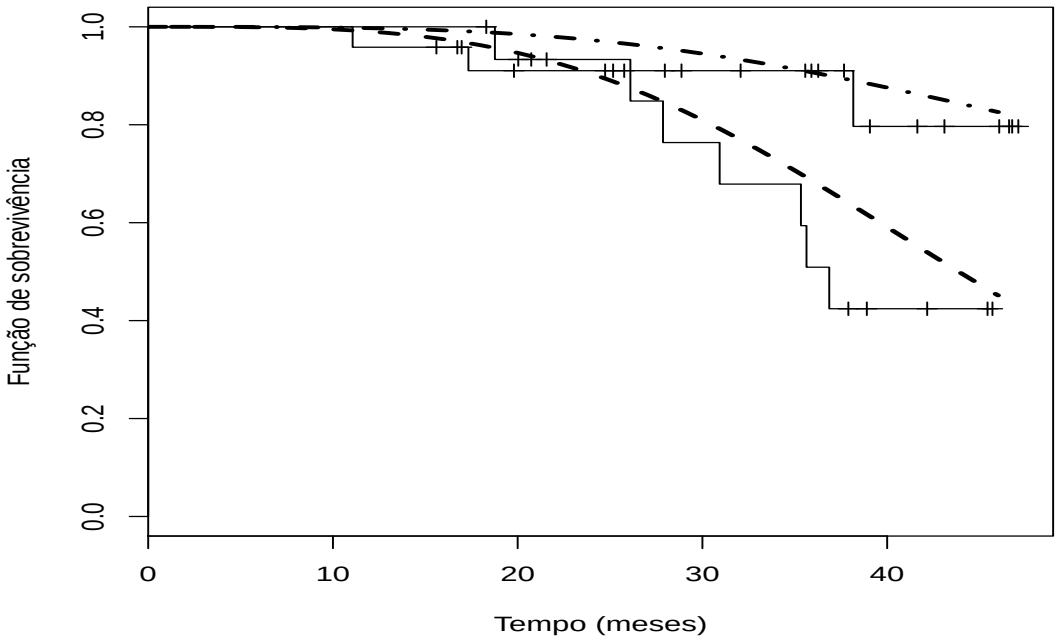

Figura 3.7: Estimativa de Kaplan-Meier e função de sobrevivência estimada pelo modelo BNGW via a inferência bayesiana: $i = 1$ (- · - · -) e $i > 1$ (- - -).

Capítulo 4

Modelo binomial negativo generalizado log-logístico

Na Seção 3.1 propomos um modelo capaz de acomodar causas competitivas latentes em que o número de causas competitivas segue uma distribuição BNG. Na Seção 3.3 abordamos este modelo especificando a distribuição Weibull para o tempo de cada causa. Outra possibilidade de distribuição para o tempo de ocorrência de cada causa, $X_l, l = 1, \ldots, m$, é a distribuição log-logística com parâmetros $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$, em que $\gamma_1 > 0$ denota o parâmetro de escala e $\gamma_2 > 0$ denota o parâmetro de forma. As funções distribuição e densidade são dadas, respectivamente, por

$$F(x|\gamma_1, \gamma_2) = \frac{x^{\gamma_2}}{\gamma_1^{\gamma_2} + x^{\gamma_2}} \quad \text{and} \quad f(x|\gamma_1, \gamma_2) = \frac{(\gamma_2/\gamma_1)(x/\gamma_1)^{\gamma_2 - 1}}{[1 + (x/\gamma_1)^{\gamma_2}]^2}. \tag{4.1}$$

Uma das vantagens da distribuição log-logística é que sua função distribuição acumulada pode ser escrita de forma fechada, o que é particularmente útil na análise de dados de sobrevivência com censura (Bennett, 1983). Outra importante característica dessa distribuição é que sua função de risco é monótona decrescente se $\gamma_2 \leq 1$ e se $\gamma_2 > 1$ é *hump-shaped*, ou seja, cresce inicialmente e decresce posteriormente. Essa situação é comum em dados de sobrevivência após cirurgia, em que o risco de infecção, hemorragia, ou outras complicações cresce e depois decresce (Kein & Moeschberger, 2003).

A inclusão de covariáveis no modelo será no parâmetro de forma da distribuição log-logística por $\gamma_{2j} = \exp(\boldsymbol{\beta}^T \boldsymbol{z}_j)$, $j = 1, \ldots, n$, em que $\boldsymbol{z}_j$ é o vetor de covariáveis para o j-ésimo indivíduo e $\boldsymbol{\beta}$ o vetor de parâmetros desconhecido sem intercepto. Logo, temos um modelo que considera o número de causas competitivas seguindo uma distribuição GNB

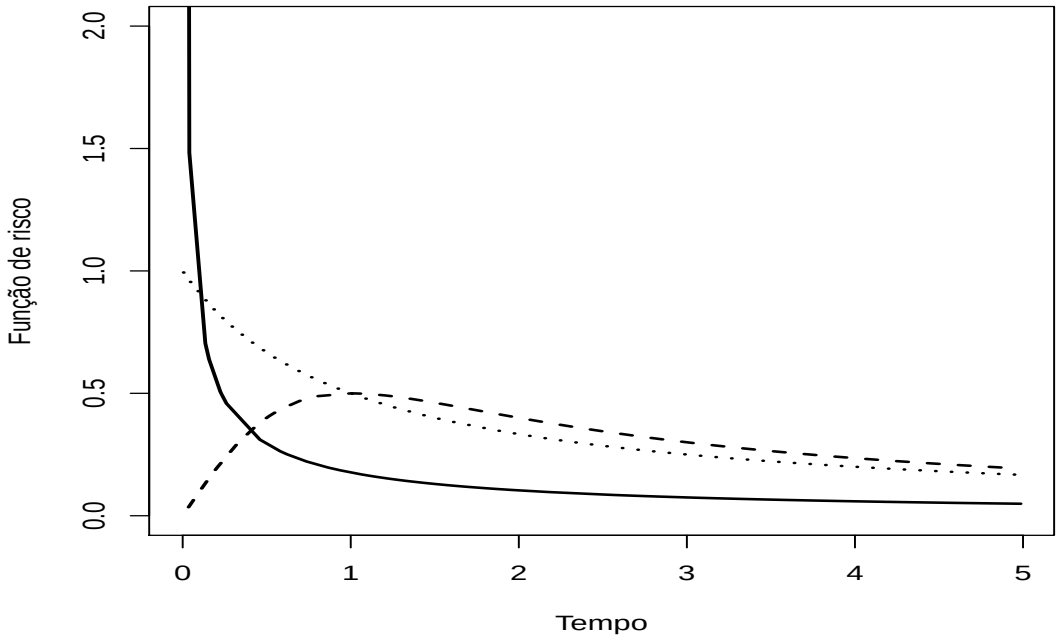

Figura 4.1: Função de risco da distribuição log-logística para $\gamma_1 = 1$ e $\gamma_2 = 0,5$ (—), $\gamma_2 = 1,0$ ($\cdots$) e $\gamma_2 = 2,0$ (- - -)

e o tempo de cada causa seguindo uma distribuição log-logística, ao qual chamaremos de modelo binomial negativo generalizado log-logístico ou abreviadamente modelo GNBLL.

Combinando as expressões (3.10) e (4.1) obtemos a função de verossimilhança do modelo BNGLL dada por

$$L(\lambda, \eta, \gamma_1, \boldsymbol{\beta}|\boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{i}) \propto \prod_{j=1}^{n} \left\{ i_j \, \frac{(\gamma_{2j}/\gamma_1)(t_j/\gamma_1)^{\gamma_{2j}-1}}{[1+(t_j/\gamma_1)^{\gamma_{2j}}]^2} \frac{ad-bc}{\left[c - d\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})\right]^2} \right\}^{\delta_j} \times \prod_{j=1}^{n} \left[\frac{a - b\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})}{c - d\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})} \right]^{i_j - \delta_j}, \quad (4.2)$$

em que $\gamma_{2j} = \exp(\boldsymbol{\beta}^T \boldsymbol{z}_j)$, $j = 1, \ldots, n$.

4.1 Estimação de máxima verossimilhança

Primeiramente propomos estimar os parâmetros maximizando a função de verossimilhança, ou seja, fazendo uso da chamada inferência clássica. Devido a complexidade da função de verossimilhança do modelo BNGLL dada em (4.2), seu ponto de máximo

não é calculado por uma expressão analítica. É necessário o uso de algum procedimento numérico como por exemplo os implementados no sistema R. A rotina *optim* é um deles e nos fornece também a matriz hessiana das estimativas.

4.1.1 Estudo de simulação

Um estudo de simulação foi conduzido para obtermos as probabilidades de cobertura dos intervalos de confiança assintóticos propostos na Seção 3.4 para pequenos e moderados tamanhos amostrais, $n = 30$, 50, 70 e 100. Sob o mesmo cenário da Seção 3.4.1, assumimos que os tempos de cada causa competitiva seguem uma distribuição log-logística com parâmetro de escala $\gamma_1 = 2,5$ e assumimos que as covariáveis são relacionadas aos tempos de vida por meio de duas covariáveis binárias, ou seja, $\gamma_{2_j} = \exp(\boldsymbol{\beta}^T \boldsymbol{z}_j) = \exp(\beta_1 z_{1_j} + \beta_2 z_{2j}), j = 1, \ldots, n$, com $\boldsymbol{\beta} = (-1,2; 2,3)$. Supomos que o número de causas competitivas segue uma distribuição BNG com $\lambda = 2,0$ e $\eta = 0,6$. Nas simulações consideramos que todos os indivíduos apresentam inicialmente o mesmo número de linfonodos contaminados, $i_j = i = 2, j = 1, \ldots, n$, sobre a hipótese de que o tratamento é dado em 5 doses mensais. Para estes valores dos parâmetros a probabilidade de cura é aproximadamente igual a 21%.

Procedimento semelhante ao descrito na Seção 3.4.1 foi utilizado para a geração dos dados. A diferença entre o esquema utilizado e o apresentado anteriormente está no segundo item, precisamente na função $F^{-1}(\cdot)$, substituído por:

2. Se $u_j < p_0$, então $y_j = \infty$. Caso contrário, $y_j = F^{-1}\left(1 - (a - cu_j^{1/i_j})/(b - du_j^{1/i_j})\right)$, em que $F^{-1}(\cdot) = \gamma_1[\cdot/(1-\cdot)]^{1/\gamma_{2j}}$.

Para cada tamanho amostral, mil simulações foram realizadas, em que a porcentagem de censura varia entre 20 e 30%. As estimativas de máxima verossimilhança assim como as probabilidades de cobertura de cada parâmetro do modelo foram calculadas como o descrito na Seção 3.4.1. Escolhemos como ponto inicial do algoritmo os verdadeiros valores dos parâmetros. As simulações que não convergiram foram descartadas. A Tabela 4.1 apresenta as PCs e permite concluirmos que seus valores convergem para o valor nominal com o aumento do tamanho amostral.

Para o modelo BNGLL também testamos a performance da estatística da razão de verossimilhanças dada em (3.15), com nível de significância nominal de 5%, na comparação

Tabela 4.1: Probabilidades de cobertura empíricas para os intervalos de confiança dos parâmetros de interesse para $n = 30$, 50, 70 e 100.

Parâmetro	Tamanho amostral			
	30	50	70	100
λ	0,972 (0,829)	0,989 (0,668)	0,966 (0,472)	0,983 (0,433)
η	0,901 (0,224)	0,928 (0,185)	0,913 (0,204)	0,950 (0,182)
γ_1	0,954 (2,121)	0,960 (1,238)	0,960 (1,679)	0,965 (1,325)
β_1	0,904 (0,567)	0,952 (0,388)	0,947 (0,353)	0,952 (0,355)
β_2	0,908 (0,807)	0,952 (0,424)	0,953 (0,552)	0,966 (0,486)

Tabela 4.2: Taxas de rejeição na comparação do modelo BNGLL contra o modelo Poisson a um nível de significância nominal de 5%.

Tamanho Amostral	$\lambda = 2$	$\lambda = 1$	$\lambda = 0,5$	$\lambda = 0,1$
30	0,998	0,680	0,102	0,025
70	0,999	0,873	0,244	0,056
100	0,999	0,951	0,275	0,027
200	0,999	0,999	0,378	0,038

do modelo BNGLL contra o modelo Poisson log-logístico. O poder do teste e o tamanho do teste foram calculados para diferentes valores do parâmetro λ e diferentes tamanhos amostrais, $n = 30$, 50, 70 e 100. A Tabela 4.2 apresenta os resultados do poder do teste que confirmam o que é esperado teoricamente: dificuldade de distinção entre um modelo e outro conforme o valor de λ diminui. A taxa de rejeição da hipótese nula fica em torno de 5% quando a hipótese alternativa converge para o modelo Poisson, o que era esperado teoricamente. Analogamente, 1000 réplicas do conjunto de dados sob o modelo Poisson foram construídas para analisar o desempenho da estatística Λ na comparação do modelo Poisson contra o modelo proposto. A taxa de rejeição da hipótese nula ficou abaixo do nível de significância de 5% para $n = 50$, 70, 100 e 200, como mostra a Tabela 4.3.

4.2 Inferência bayesiana

As distribuições *a priori* dos parâmetros foram escolhidas de acordo com o espaço paramétrico de cada um deles, o que significa que $\lambda \sim$ Log-normal(a_0, a_1), $\eta \sim \mathcal{B}(b_0, b_1)$,

Tabela 4.3: Taxa de rejeição da hipótese nula na comparação do modelo Poisson contra o modelo proposto BNGLL.

$n = 30$	$n = 50$	$n = 100$	$n = 200$
0,001	0,001	0,001	0,039

$\gamma_1 \sim \Gamma(c_0, c_1)$ e que as G componentes de $\boldsymbol{\beta}$ são independentes *a priori* e cada β_g tem distribuição a normal, $\mathcal{N}(\mu_{\beta_g}, \sigma^2_{\beta_g})$, em que $a_0, a_1, b_0, b_1, c_0, c_1, \mu_{\beta_g}, \sigma^2_{\beta_g}$, $g = 1, \ldots, G$, são hiperparâmetros conhecidos.

$$\pi(\lambda, \eta, \gamma_1, \boldsymbol{\beta}|\boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{i})) \propto \prod_{j=1}^{n} \left\{ i_j \, \frac{(\gamma_{2j}/\gamma_1)(t_j/\gamma_1)^{\gamma_{2j}-1}}{[1+(t_j/\gamma_1)^{\gamma_{2j}}]^2} \frac{ad-bc}{\left[c - d\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})\right]^2} \right\}^{\delta_j}$$
$$\times \prod_{j=1}^{n} \left[\frac{a - b\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})}{c - d\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})} \right]^{i_j - \delta_j} \pi(\lambda|a_0, a_1)\pi(\eta|b_0, b_1)\pi(\gamma_1|c_0, c_1) \prod_{g=1}^{G} \pi(\beta_g|\mu_g, \sigma_g^2)$$

Independentemente das distribuições *a priori* escolhidas, a distribuição *a posteriori* do modelo proposto é analiticamente intratável. Como alternativa usamos os métodos de Monte Carlo em Cadeias de Markov (MCMC) com o amostrador de Gibbs ou o algoritmo de Metropolis-Hastings (veja p. ex. Chib & Greenberg, 1995). As distribuições condicionais completas dos parâmetros são dadas a seguir:

$$\pi(\lambda|\eta, \gamma_1, \boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{i})) \propto \prod_{j=1}^{n} \frac{[ad-bc]^{\delta_j}[a - b\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})]^{i_j-\delta_j}}{[c - d\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})]^{i_j+\delta_j}} \pi(\lambda|a_0, a_1)$$

$$\pi(\eta|\lambda, \gamma_1, \boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{i})) \propto \prod_{j=1}^{n} \frac{[ad-bc]^{\delta_j}[a - b\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})]^{i_j-\delta_j}}{[c - d\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})]^{i_j+\delta_j}} \pi(\eta|b_0, b_1)$$

$$\pi(\gamma_1|\lambda, \eta, \boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{i})) \propto \prod_{j=1}^{n} \frac{[(\gamma_{2j}/\gamma_1)(t_j/\gamma_1)^{\gamma_{2j}-1}]^{\delta_j}[a - b\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})]^{i_j-\delta_j}}{[1+(t_j/\gamma_1)^{\gamma_{2j}}]^{2\delta_j}[c - d\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})]^{i_j+\delta_j}} \pi(\gamma_1|c_0, c_1)$$

e

$$\pi(\beta_g|\lambda, \eta, \gamma_1, \boldsymbol{\beta}_{-g}, \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{i})) \propto \prod_{j=1}^{n} \frac{[\gamma_{2j}(t_j/\gamma_1)^{\gamma_{2j}-1}]^{\delta_j}[a - b\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})]^{i_j-\delta_j}}{[1+(t_j/\gamma_1)^{\gamma_{2j}}]^2 \left[c - d\gamma_{2j}/(\gamma_1^{\gamma_{2j}} + t_j^{\gamma_{2j}})\right]^{i_j-\delta_j}} \pi(\beta_g|\mu_g, \sigma_g^2),$$

em que $\boldsymbol{\beta}_{-g} = (\beta_1, \ldots, \beta_{g-1}, \beta_{g+1}, \ldots, \beta_G)$, ou seja, o vetor de parâmetros $\boldsymbol{\beta}$ sem a g-ésima componente. Devido a todas as distribuições condicionais *a posteriori* não serem obtidas analiticamente, faremos uso do algoritmo de Metropolis-Hastings.

4.2.1 Estudo de simulação

Com os mesmos objetivos do estudo descrito na Seção 3.5.1 e de maneira análoga, realizamos um estudo de simulação assumindo que os pacientes estão em tratamento contra o

Tabela 4.4: Probabilidades de cobertura empíricas para os intervalos de credibilidade dos parâmetros de interesse e amplitude média dos intervalos de credibilidade (entre parênteses) para $n = 30$, 50, 70 e 100.

Parâmetro	Tamanho amostral			
	30	50	70	100
λ	0,987 (1,66)	0,980 (1,55)	0,977 (1,59)	0,982 (1,51)
η	0,981 (0,35)	0,976 (0,31)	0,981 (0,29)	0,974 (0,27)
γ_1	0,982 (4,66)	0,987 (3,02)	0,986 (2,84)	0,990 (2,59)
β_1	0,993 (0,86)	0,982 (0,80)	0,985 (0,79)	0,981 (0,76)
β_2	0,990 (1,53)	0,981 (1,56)	0,983 (1,51)	0,987 (1,63)

câncer, dado em 5 doses com intervalo entre uma dose e outra igual a um mês e todos inicialmente apresentam dois linfonodos contaminados, $i_j = i = 2, j = 1, \ldots, n$. Os tempos de cada causa seguem uma distribuição log-logística com parâmetros γ_1 e γ_2, sendo $\gamma_1 = 2,5$ e γ_2 relacionado a duas covariáveis através de $\gamma_{2_j} = \exp\{\boldsymbol{\beta}^T \boldsymbol{z}_j\} = \exp\{\beta_1 z_{1_j} + \beta_2 z_{2j}\}, j = 1, \ldots, n,$, com $\boldsymbol{\beta} = (-1,2; 2,3)$. A distribuição BNG para o número de causas competitivas com parâmetros $\lambda = 1,8$ e $\eta = 0,4$ completa nossas suposições. Nestas condições a probabilidade de cura é de $51,3\%$. Amostras de tamanho $n = 30$, 50, 70 e 100 foram obtidas como no esquema apresentado na Seção 4.1.1.

Utilizamos distribuições *a priori* independentes para os parâmetros com variâncias grandes afim de assegurar que sejam não informativas: $\lambda \sim$ Log-normal$(0; 10)$, $\eta \sim \mathcal{B}(1,1)$, $\gamma_1 \sim \Gamma(1; 0,001)$, β_1 e β_2 têm distribuição *a priori* normal, $\mathcal{N}(0; 1000)$. Construímos, para cada parâmetro, duas cadeias paralelas de tamanho amostral 11000. As primeiras 1000 amostras foram selecionadas como sendo o período de *burn-in* e as restantes selecionadas de 5 em 5, resultando numa amostra de tamanho 4000. Procedendo dessa forma, 1000 simulações foram feias para cada valor de n. As rotinas foram implementadas no sistema R, usando o pacote BRugs. Na Tabela 4.4 são apresentadas as PCs e as amplitudes médias dos intervalos de credibilidade 95% para cada parâmetro. Tais resultados nos permitem concluir que tanto as PCs como as amplitudes médias não sofrem variação significativa com o aumento do tamanho amostral.

Tabela 4.5: Verdadeiros valores, estimativas e intervalos de confiança obtidos pelas estimativas de máxima verossimilhança.

Parâmetro	Verdadeiro valor	Estimativa	DP	IC 95%
λ	1,8	1,551	0,300	(0,862; 2,039)
η	0,4	0,458	0,072	(0,317; 0,599)
γ_1	2,5	2,418	0,331	(1,771; 3,066)
β_1	-1,2	-1,409	0,230	(-1,764;-1,054)
β_2	2,3	2,281	0,181	(1,828; 2,732)

4.3 Dados artificiais

Dados artificiais foram construídos sob um cenário pós-cirúrgico de pacientes com certo tipo de câncer. Imediatamente após a intervenção cirúrgica para a retirada dos tumores e da região contaminada (linfonodos contaminados) o risco de os pacientes virem a óbito aumenta e posteriormente diminui. Os pacientes iniciam a segunda etapa do tratamento quimioterápico logo após a cirurgia, visando, entre outros, a descontaminação dos linfonodos infectados não retirados no procedimento cirúrgico. Uma amostra de tamanho 70 com os mesmos valores dos parâmetros atribuídos na Seção 4.1.1 foi construída afim de compararmos as estimativas a seus valores verdadeiros nos dois tipos de inferências propostos.

Procedendo como na Seção 4.1.1 obtivemos as estimativas de máxima verossimilhança dos parâmetros do modelo BNGLL que são, juntamente com o desvio padrão e o intervalo de confiança 95%, apresentados na Tabela 4.5. A probabilidade de cura estimada, $\widehat{p_0}$, é igual a $49,1\%$. O desvio padrão de cada estimativa foi calculado através da matriz hessiana estimada.

Sob as mesmas considerações da Seção 4.2.1 obtivemos as médias *a posteriori*, o desvio padrão e os intervalos de credibilidade 95% para cada parâmetro do modelo descrito na Tabela 4.6. Tais estimativas nos permitem concluir que a probabilidade de cura estimada é igual a 46,8%. O gráfico do ajuste de Kaplan-Meier juntamente com a função de sobrevivência estimada nos casos clássico e bayesiano são apresentados na Figura 4.2

Comparando os resultados das Tabelas 4.5 e 4.6 concluímos que não há diferença significativa nas estimativas dos parâmetros obtidas pela maximização da verossimilhança e

Tabela 4.6: Verdadeiros valores, médias *a posteriori* e intervalos de credibilidade obtidos pela inferência bayesiana.

Parâmetro	Verdadeiro valor	Média *a posteriori*	DP	ICred 95%
λ	1,8	1,535	0,341	(0,498; 1,960)
η	0,4	0,487	0,079	(0,345; 0,661)
γ_1	2,5	2,520	0,496	(1,800; 3,759)
β_1	-1,2	-1,471	0,221	(-2,002; -1,107)
β_2	2,3	2,164	0,390	(1,612; 2,644)

pela inferência bayesiana, resultando na proximidade entre as probabilidades de cura estimadas. No entanto, concluímos que os desvios padrão da inferência clássica são menores, implicando em intervalos de confiança de menores amplitudes, comparadas às amplitudes dos intervalos de credibilidade.

4.4 Dados de melanoma cutâneo

Os dados foram captados de Ibrahim *et al.* (2001) (veja também Kirkwood, Ibrahim, Sondak, Richards, Flaherty, Ernstoff, Smith, Rao, Steele & Blum, 2000) e apresentam os tempos de sobrevivência (em anos) provenientes de um estudo de melanoma cutâneo que acompanhou 427 pacientes. Analisamos os pacientes classificados por 3 e 4 nódulos que totalizam 169 pacientes, dos quais 87 estão classificados por 3 nódulos e 82 por 4 nódulos. Compõem o conjunto de dados as seguintes informações: idade (em anos, média 47,7 e desvio padrão 12,3); sexo (0: masculino, $n = 64$ e 1: feminino, $n = 105$); medida *Breslow*, uma medida de quão invasivo é o melanoma, (em milímetros, média 3,5 e desvio padrão 3,4); performance, uma medida de bem-estar do paciente frente a suas atividades diárias (0: ativo, $n = 142$ e 1: outra, $n = 27$); tratamento (0: observação, $n = 83$ e 1: tratamento, $n = 86$). A porcentagem de dados censurados é 43,2%. A variável categórica nódulos foi utilizada como a quantidade inicial de linfonodos contaminados. Como é comum, supomos que o tratamento é dado em 4 doses mensais, ou seja, $k = 4$ e $\tau = 1/12$.

Para identificar o comportamento da função de risco dos tempos observados, construímos seu gráfico TTT descrito na Seção 3.7. A Figura 4.3 apresenta o gráfico TTT dos dados de melanoma cutâneo para 3 e 4 nódulos, que indica um crescimento inicial do

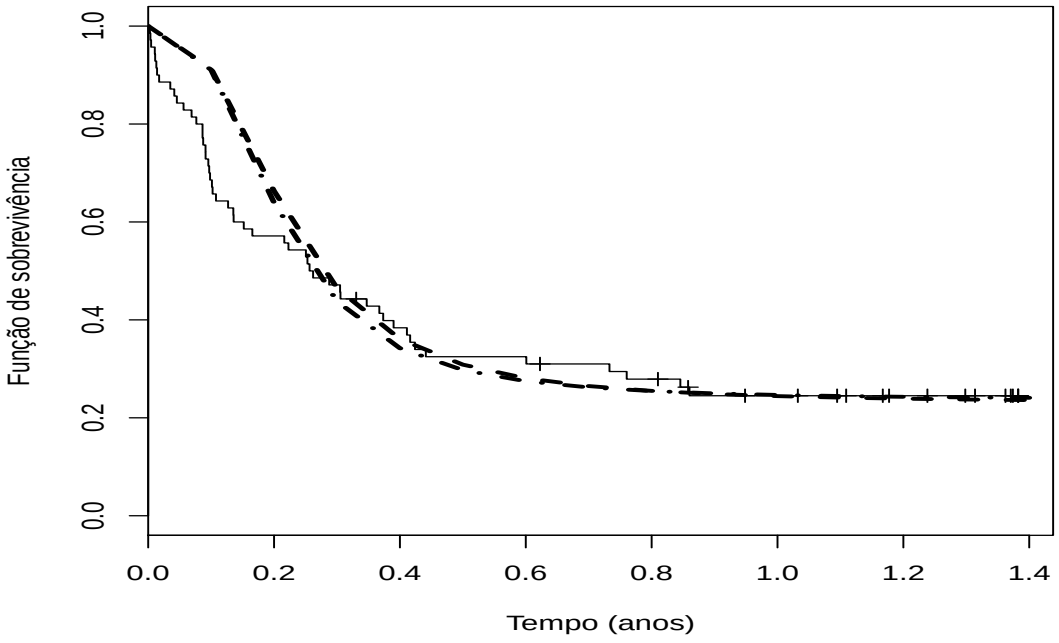

Figura 4.2: Estimativa de Kaplan-Meier e função de sobrevivência estimada pelo modelo BNGLL via inferência clássica (- · - · -) e via inferência bayesiana (- - -) - dados artificiais.

risco que permanece constante posteriormente. Tal situação pode ser representada pela distribuição log-logística.

Ajustamos o modelo BNGLL abordando como covariáveis as cinco informações descritas anteriormente, ou seja, idade, sexo, tratamento, medida *Breslow* e performance. Quatro modelos foram ajustados: Poisson log-logístico (PLL), binomial negativo log-logístico (BNLL), BNGW e BNGLL. A Tabela 4.7 apresenta os valores de máximo da log-verossimilhança, $\max \log L(\cdot)$, e os valores das estatísticas AIC e BIC para os quatro modelos ajustados. Comparando essas estatísticas, notamos que as diferenças são pequenas entre os modelos PLL, BNLL e BNGLL, embora evidenciam a favor do modelo BNGLL. No entanto o modelo BNGLL permite melhor interpretação e adaptação ao estudo estimando, de certa forma, a eficácia de cada dose do tratamento e a taxa do processo de metástase. Vale ressaltar que os valores das estatísticas AIC e BIC do modelo BNGW, na comparação com os demais modelos, o classifica como o menos adequado.

Os resultados das estimativas de máxima verossimilhança dos parâmetros do modelo BNGLL, seus desvios padrão e seus intervalos de confiança 95% são apresentados na Tabela 4.8 e mostram que apenas a covariável medida *Breslow* é significativa. Os resultados

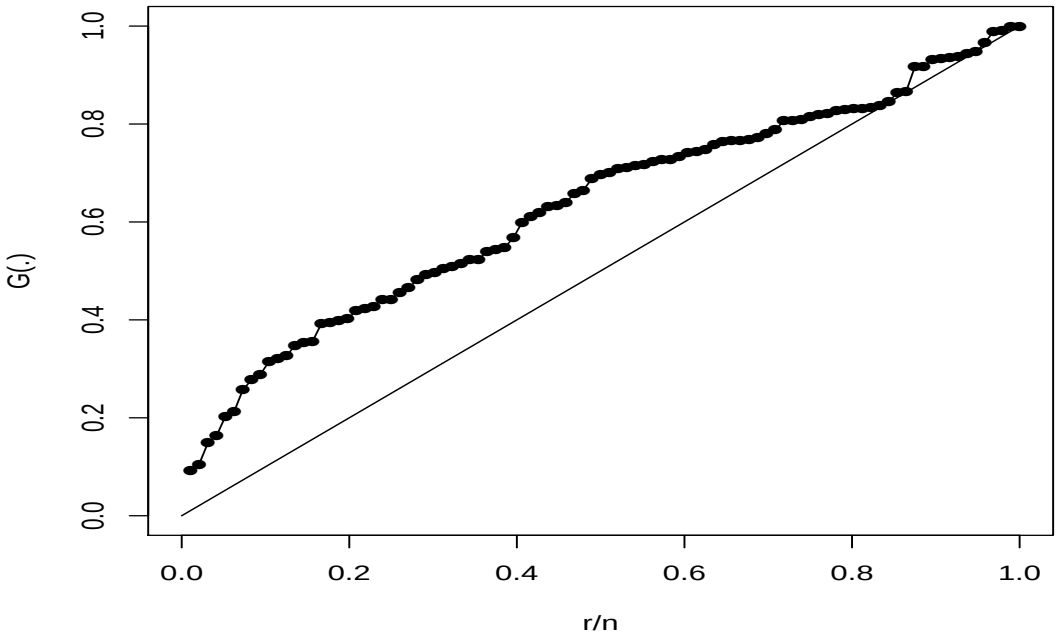

Figura 4.3: Gráfico TTTplot dos dados de melanoma cutâneo, considerando 3 e 4 nódulos.

Tabela 4.7: Valores $\max \log L(\cdot)$ e estatísticas AIC e BIC para os três modelos ajustados, PLL, BNLL e BNGLL, considerando as cinco covariáveis.

Critério	PLL	BNLL	BNGW	BNGLL
$\max \log L(\cdot)$	-231,63	-231,43	-236,27	-228,63
AIC	477,26	476,86	488,55	473,27
BIC	499,16	498,77	513,58	498,30

do ajuste que considera apenas a covariável significativa são apresentados nas Tabelas 4.9 e 4.10 e novamente dão evidências a favor do modelo BNGLL. A probabilidade de cura estimada para o modelo selecionado, $\widehat{p_0}$, para a categoria 3 nódulos é igual a 42,0% e para a categoria 4 nódulos é 31,5%. A Figura 4.4 ilustra o ajuste de Kaplan-Meier separados por quantidade de nódulos, juntamente com a função de sobrevivência estimada pelo modelo BNGLL.

Os ajustes para os modelos PLL, BNLL, BNGW e BNGLL também foram obtidos pela inferência bayesiana. Escolhemos distribuições *a priori* independentes e não informativas, tais que $\eta \sim \mathcal{B}(1,1)$, $\gamma_1 \sim \Gamma(1;0,01)$ e os βs têm distribuição normal $\mathcal{N}(0;100)$. Os problemas de convergência foram burlados considerando uma reparametrização logarítmica

Tabela 4.8: Estimativas de máxima verossimilhança dos parâmetros do modelo BNGLL, desvios padrão e intervalos de confiança 95% (IC 95%), considerando as cinco covariáveis.

Parâmetro	Estimativa	DP	IC 95%
λ	0,415	0,971	(0,062; 2,786)
η	0,724	0,058	(0,701; 0,747)
γ_1	1,615	0,132	(1,245; 2,095)
β_{sexo}	-0,008	0,194	(-0,389; 0,373)
$\beta_{\text{performance}}$	-0,131	0,294	(-0,709; 0,447)
β_{idade}	0,006	0,004	(-0,002; 0,013)
β_{Breslow}	0,149	0,039	(0,072; 0,225)
$\beta_{\text{tratamento}}$	0,119	0,226	(-0,323; 0,562)

Tabela 4.9: Valores $\max \log L(\cdot)$ e estatísticas AIC e BIC para os três modelos ajustados, PLL, BNLL e BNGLL, considerando apenas a covariável significativa.

Critério	PLL	BNLL	BNGW	BNGLL
$\max \log L(\cdot)$	-235,02	-234,32	-253,99	-231,69
AIC	476,04	474,66	515,97	472,38
BIC	485,43	484,05	528,49	483,90

para o parâmetro lambda, tal que $\log(\lambda) \sim \mathcal{N}(-5; 0, 4)$. O código computacional foi implementado no OpenBUGS versão 3 (Spiegelhalter *et al.*, 1999) e os gráficos foram construídos com o auxílio do sistema R (R Development Core Team, 2009). Duas cadeias paralelas de tamanho 60000 foram geradas para cada parâmetro. As primeiras 10000 foram descartadas e as restantes selecionadas de 20 em 20, totalizando uma amostra de tamanho 5000. Em uma das cadeias tomamos as estimativas de máxima verossimilhança como ponto inicial e na outra tomamos 0,5 para λ e para η, 1 para γ_1 e para 0 os βs. A convergência das cadeias foi monitorada pela análise gráfica de Geweke (Geweke, 1992) e pelo método Gelman-Rubin (Brooks & Gelman, 1998) calculado pelo OpenBUGS. Para verificar a sensibilidade aos hiperparâmetros de dispersão, consideramos outros valores e os resultados *a posteriori* foram similares.

Idade, sexo, medida *Breslow*, tratamento e performance foram as covariáveis incorporadas na análise. As estatísticas DIC foram obtidas para os modelos ajustados PLL, BNLL, BNGW e BNGLL, sendo seus respectivos valores iguais a 477,7, 477,0, 487,4 e

Tabela 4.10: Estimativas de máxima verossimilhança dos parâmetros do modelo BNGLL, desvios padrão e intervalos de confiança 95% (IC 95%), considerando apenas a covariável significativa.

Parâmetro	Estimativa	DP	IC 95%
λ	0,006	0,098	(0,005;0,007)
η	0,708	0,017	(0,700;0,714)
γ_1	1,647	0,075	(1,421;1,908)
β_{Breslow}	0,204	0,021	(0,163;0,245)

471,9, o que favorece o modelo BNGLL. Os resultados contidos na Tabela 4.11, médias *a posteriori*, desvios padrão e intervalos de credibilidade, indicam que apenas a covariável medida *Breslow* é significativa, o que corrobora com as estatísticas anteriores.

Fizemos outro ajuste considerando apenas a covariável significativa. Nessa condição, os valores da estatística DIC dos modelos PLL, BNLL, BNGW e BNGLL são respectivamente iguais a 476,3, 475,0, 515,8 e 471,3, dando evidências a favor do modelo BNGLL. A Tabela 4.12 apresenta as médias *a posteriori* das amostras selecionadas para cada parâmetro, assim como os desvios padrão e os intervalos de credibilidade 95% das estimativas. As probabilidades de cura estimadas para 3 e 4 nódulos são iguais a 41,7% e 31,2%, respectivamente. As Figuras 4.5 e 4.6 apresentam, respectivamente, o histórico das cadeias das amostras selecionadas e as densidades marginais *a posteriori* aproximadas para cada parâmetro. Os gráficos do ajuste de Kaplan-Meier e da sobrevivência estimada pelo modelo selecionado no caso bayesiano estão na Figura 4.7.

Da análise realizada provêm algumas informações interessantes. Particularmente, λ denota a taxa de proliferação da contaminação dos linfonodos, que é praticamente nula, indicando o controle da doença, ou seja, a não ocorrência do processo de metástase. O parâmetro η denota a probabilidade de contaminação do linfonodo depois de cada dose do tratamento, igual a quase 70%. Assumindo doses independentes, a probabilidade de a contaminação permanecer no linfonodo após as 4 doses do tratamento é igual a $\eta^4 = 24\%$, ou seja, é de 76% a probabilidade de descontaminação do linfonodo ao término do tratamento. Ainda temos que $\gamma_{2j} > 1$ para todos os pacientes, $j = 1, \ldots, 169$, implicando que o risco cresce inicialmente e posteriormente decresce, confirmando a informação do gráfico TTT.

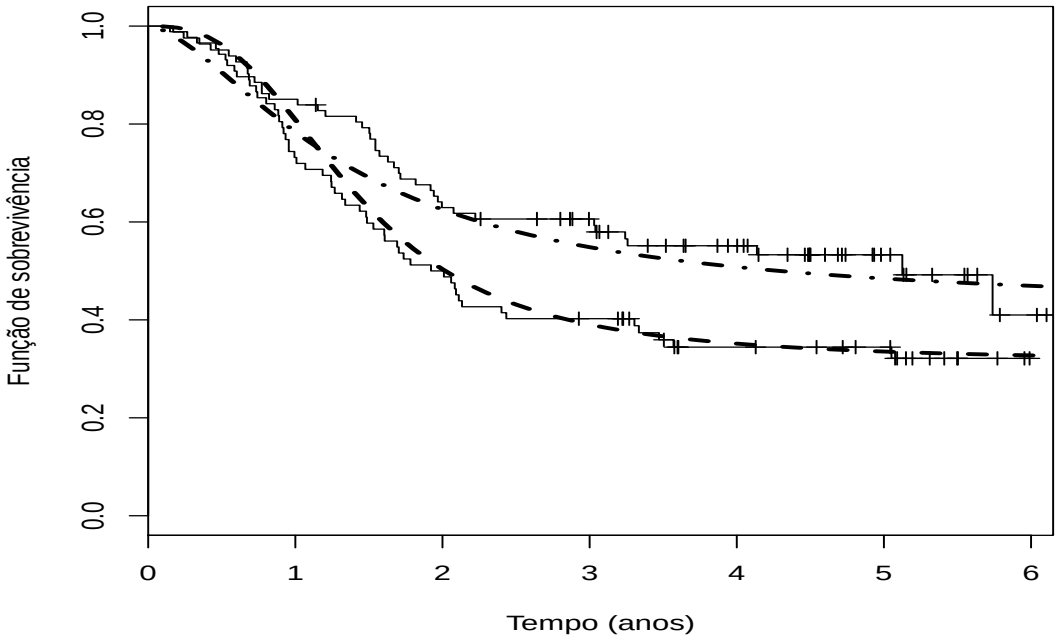

Figura 4.4: Estimativa de Kaplan-Meier e função de sobrevivência estimada pelo modelo BNGLL via inferência clássica: $i = 3$ (- · - · -) e $i = 4$ (- - -).

Os resultados obtidos pela estimação de máxima verossimilhança e pela inferência bayesiana são próximos e implicam nas mesmas conclusões a respeito do modelo a ser escolhido e das covariáveis a serem consideradas.

4.5 Comentários finais

Os modelos BNGW e BNGLL acomodam dados de sobrevivência na presença de causas competitivas latentes. A principal diferença entre eles é que o primeiro modela dados com função de risco crescente, decrescente e constantes. Já o segundo acomoda dados cuja função de risco pode primeiramente crescer e posteriormente decrescer. Ambas as situações são encontradas em estudos clínicos, mais precisamente em estudos da área oncológica. Novamente destacamos que a principal vantagem dos dois modelos é a estimação da taxa de espalhamento da doença e a eficácia de cada dose do tratamento.

A inferência clássica e a inferência bayesiana apresentaram resultados similares. Futuramente informações *a priori* mais particulares poderão ser fornecidas por especialistas e incorporadas à análise.

Tabela 4.11: Média *a posteriori* dos parâmetros do modelo BNGLL, desvios padrão e intervalos de credibilidade 95% (ICred 95%), considerando as cinco covariáveis.

Parâmetro	Média *a posteriori*	DP	ICred 95%
λ	0,017	0,037	(0,001; 0,110)
η	0,706	0,0172	(0,672; 0,741)
γ_1	1,935	0,374	(1,542; 3,005)
β_{sexo}	-0,011	0,184	(-0,383; 0,335)
$\beta_{\text{performance}}$	-0,076	0,294	(-0,656; 0,495)
β_{idade}	0,006	0,004	(-0,002; 0,013)
β_{Breslow}	0,121	0,044	(0,027; 0,201)
$\beta_{\text{tratamento}}$	0,176	0,220	(-0,251; 0,611)

Tabela 4.12: Média *a posteriori* dos parâmetros do modelo BNGLL, desvios padrão e intervalos de credibilidade 95% (ICred 95%), considerando apenas a covariável significativa.

Parâmetro	Média *a posteriori*	DP	ICred 95%
λ	0,014	0,036	(0,001;0,095)
η	0,709	0,018	(0,674;0,744)
γ_1	1,683	0,120	(1,521;1,953)
β_{Breslow}	0,195	0,024	(0,142;0,237)

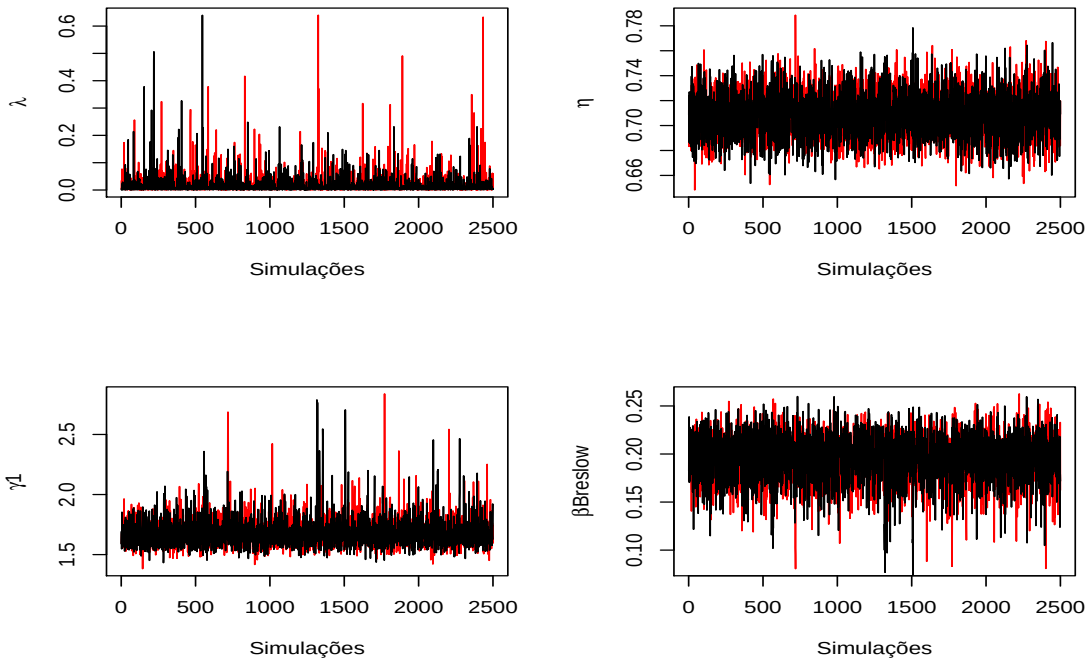

Figura 4.5: Histórico das cadeias.

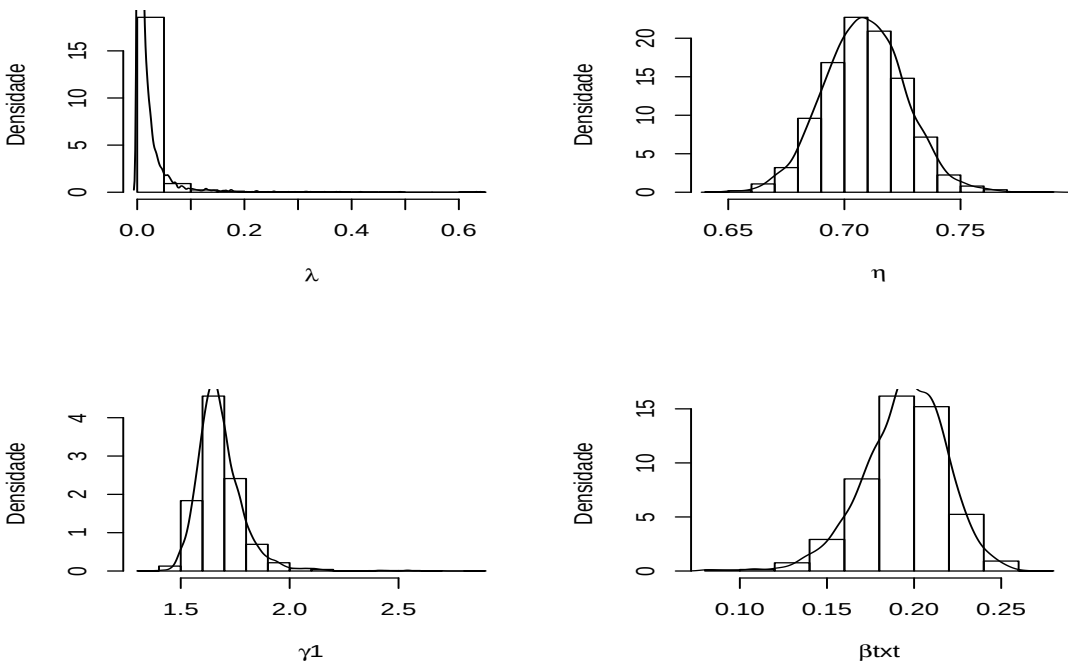

Figura 4.6: Densidades *a posteriori* marginais aproximada dos parâmetros.

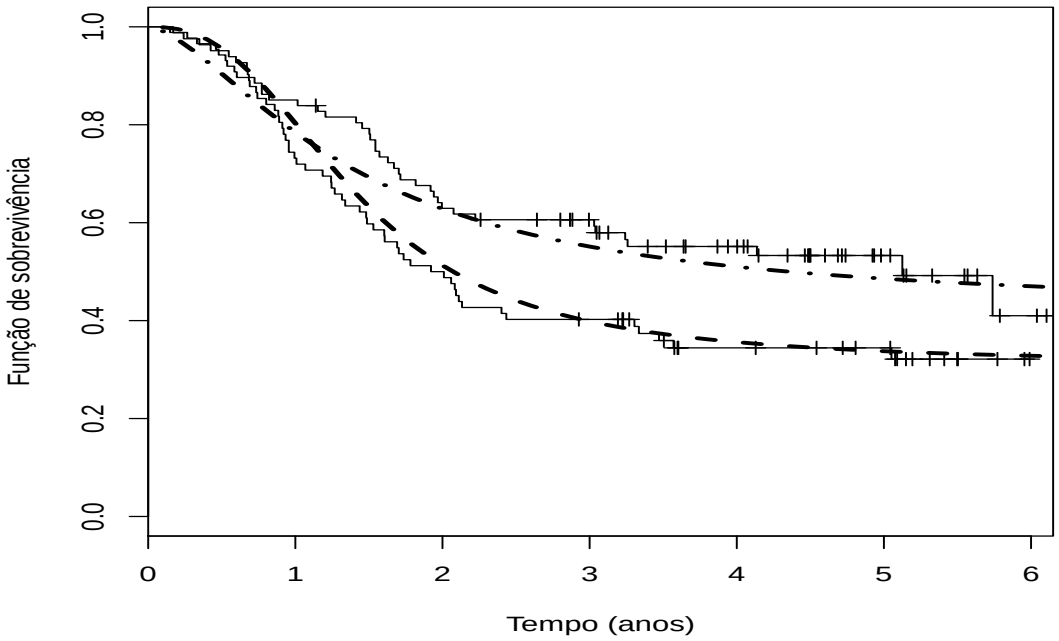

Figura 4.7: Estimativa de Kaplan-Meier e função de sobrevivência estimada pelo modelo BNGLL via a inferência bayesiana: $i = 3$ (- · - · -) e $i = 4$ (- - -).

Referências Bibliográficas

Aarset, M. V. (1985). The null distribution for a test of constant versus "bathtub" failure rate. *Scandinavian Journal of Statistics*, **12**(1), 55–68.

Balka, J., Desmond, A. F. & McNicholas, P. D. (2009). Review and implementation of cure models based on first hitting times for Wiener processes. *Lifetime Data Anal*, **15**(2), 147–176.

Banerjee, S. & Carlin, B. P. (2004). Parametric spatial cure rate model for interval-cesnored time-to-relapse data. *Biometrics*, **60**(1), 268–275.

Bennett, S. (1983). Log-logistic regression models for survival data. *Applied Statistics*, **32**(2), 165–171.

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **42**, 501–515.

Boag, J. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer theraphy. *Journal of the Royal Statistics Society, Series B*, **11**, 15–44.

Brasil (2010). Inca lança estimativa 2010: Incidência de câncer no brasil. Home page. acessado em 14 de janeiro de 2010.

Brooks, S. P. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Chen, M. H., Ibrahim, J. G. & Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association: Theory and Methods*, **94**(447), 909–919.

Chen, M. H., Ibrahim, J. G. & Sinha, D. (2002). Bayesian inference for multivariate survival data with a cure fraction. *Journal of Multivariate Analysis*, **80**(1), 101–126.

Chib, S. & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**(4), 327–335.

Cobre, J. & Louzada Neto, F. (2009). A sampling-based approach for a hybrid scale intensity model. *Journal of Statistics: Advances in Theory and Applications*, **1**(2), 159–168.

Cobre, J., Louzada Neto, F. & Perdoná, G. S. C. (2009). A generalized negative binomial Weibull distribution for survival data in presence of latent competing causes and cure fraction. Relatório Técnico do DEs - Teoria & Métodos 200, São Carlos, Brasil. ISSN 0104-0499.

Cobre, J., Louzada Neto, F. & Perdoná, G. S. C. (2010). A Bayesian analysis for the generalized negative binomial Weibull cure fraction survival model: Estimating the lymph nodes metastasis rates. Relatório Técnico do DEs - Teoria & Métodos 211, São Carlos, Brasil. ISSN 0104-0499.

Colosimo, E. A. & Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*. Edgard Blüncher, São Paulo.

Cooner, F., Banerjee, S., Carlin, B. & Sinha, D. (2007). Flexible cure rate modelling under latent activation schemes. *Journal American Statistics Association*, **102**(478), 560–572.

Cox, D. R. (1972). The statistical analysis of dependencies in point process. *Stochastic Point Processes*, pages 55–66.

Cox, D. R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

de Castro, M., Cancho, V. G. & Rodrigues, J. (2007). A flexible model for survival data with a surviving fraction. Relatório Técnico do DEs - Teoria & Métodos 173, São Carlos, Brasil.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long term survivors. *Biometrics*, **38**(4), 1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, **14**(3), 257–262.

Gail, M. H., Santner, T. J. & Brown, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics*, **36**(2), 255–266.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the America Statistical Association*, **85**(410), 398–409.

Gelman, A. & Rubin, B. D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**(4), 457–511.

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

Ghitany, M. & Maller, R. (1992). Asymptotic results for exponential mixture models with long-term survivors. *Statistics*, **23**(4), 321–336.

Ghitany, M., Maller, R. & Zhou, S. (1994). Exponential mixture models with long-term survivors and covariates. *Journal of Multivariate Analysis*, **49**(2), 218–241.

Goldman, A. I. (1984). Survivorship analysis when cure is a possibility: A monte carlo study. *Statistics in Medicine*, **3**(2), 153–163.

Gordon, N. H. (1990). Application of the theory of finite mixtures for the estimation of 'cure' rates of treated cancer pacients. *Statistics in Medicine*, **9**(4), 397–407.

Gozzo, T. O. (2008). *Toxicidade ao tratamento quimioterápico em mulheres com câncer de mama*. Tese de doutorado, Escola de Enfermagem de Ribeirão Preto, USP, Ribeirão Preto, Brasil.

Hanin, L. G. (2001). Iterated birth and death process as a model of radiation cell survival. *Mathematical Biosciences*, **169**(1), 89–107.

Hanin, L. G., Zaider, M. & Yalovlev, A. Y. (2001). Distribution of the number of clonogens surviving fractionated radiotherapy: A long-standing problem revisited. *International Journal of Radiation Biology*, **77**(2), 205–213.

Ibrahim, J. G., Chen, M. H. & Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, New York.

Kein, J. P. & Moeschberger, M. L. (2003). *Survival Analysis: Thechniques for Censored and Tuncated Data*. Springer - Verlang, New York, second edition.

Kirkwood, J. M., Ibrahim, J. G., Sondak, V. K., Richards, J., Flaherty, L. E., Ernstoff, M. S., Smith, T. J., Rao, U., Steele, M. & Blum, R. H. (2000). High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of intergroup trial E1690/S91111/C9190. *Journal of Clinica Oncology*, **18**(12), 2444–2458.

Kuk, A. Y. C. & Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, **79**(3), 531–541.

Lawless, J. & Thiagarajah, K. (1996). A point-process model incorporating renewals and time trends, with application to repairable systems. *Technometrics*, **38**(2), 131–138.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

Lawless, J. F. (1995). The analysis of recurrent events for multiple subjects. *Applied Statistics*, **44**(4), 487–498.

Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

Lawless, J. F. & Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics*, **37**(2), 158–168.

Louzada Neto, F. (2004). A hybrid scale intensity model for recurrent event dada. *Communication in Statistics*, **33**(1), 119–133.

Louzada Neto, F. (2008). Intensity models for parametric analysis of recurrente events data. *Brazilian Journal of Probability and Statistics*, **22**(1), 23–33.

Louzada Neto, F. & Cobre, J. (2008). A multiple time scale survival model with a cure fraction. Relatório Técnico do DEs - Teoria & Métodos 195, São Carlos, Brasil. ISSN 0104-0499.

Louzada Neto, F. & Cobre, J. (2010). A multiple time scale survival model. *Advances and Applications in Statistics*, **14**(1), 1–16.

Louzada Neto, F., Cobre, J. & Perdoná, G. S. C. (2009). Generalized negative binomial log-logistic cure survival rate model: An application to a cutaneous melanoma data. Relatório Técnico do DEs - Teoria & Métodos 209, São Carlos, Brasil. ISSN 0104-0499.

Maller, R. & Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, Chichester.

Massey, W. A., Parker, G. A. & Whitt, W. (1996). Estimating the parameters of a nonhomogeneous poisson process with linear rate. *Telecommunication Systems*, **5**(2), 361–388.

McDonald, J. W. & Rosina, A. (2001). Mixture modelling of recurrent events times with long-term survivors: Analysis of Hutterite birth intervals. *Statistical Methods & Applications*, **10**(3), 257–272.

McShane, B., Adrian, M., Bradlow, E. & Fader, P. (2008). Count models based on Weibull interarrival times. *Journal of Business and Economic Statistics*, **26**(3), 369–378.

Nelson, W. (1988). Graphical analysis of system repair data. *Journal of Quality Technology*, **20**(1), 24–35.

Nelson, W. (1995). Confidence limits for recurrent data-applied to cost or number of product repairs. *Technometrics*, **37**, 147–157.

Ng, S. K., McLachlan, G., Yau, K. & Lee, A. (2004). Modelling the distribution of ischaemic stroke-specific survival time using an em-based mixture approach with random effects adjustment. *Statistics in Medicine*, **23**(17), 2729–2744.

Paulino, C. D., Turkman, M. A. A. & Murteira, B. (2003). *Estatísitca Bayesiana*. Fundação Calouste Gulbenkian, Lisboa.

Peng, Y. W., Dear, K. B. G. & Denham, J. W. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, **17**(8), 813–830.

Pepe, M. S. & Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association: Theory and Methods*, **88**(423), 811–820.

Pollock, R. E., Doroshow, J. H., Khayat, D., Nakao, A. & O'Sullivan, B. (2004). *UICC Manual of Clinical Oncology, 8th Edition*. John Wiley & Sons, New York.

Prentice, R. L., Willians, B. J. & Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, **68**(2), 373–379.

Rodrigues, J., Cancho, V. G. & de Castro, M. (2008). *Teoria Unificada de Análise de Sobrevivência*. ABE, São Paulo.

Rodrigues, J., Cancho, V. G., de Castro, M. & Louzada-Neto, F. (2009a). On the unification of the long-term survival models. *Statistics & Probability Letters*, **79**(6), 753–759.

Rodrigues, J., de Castro, M., Cancho, V. G. & Balakrishnan, N. (2009b). COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, **139**(10), 3605–3611.

Sparano, J. A., Wang, M., Martino, S., Jones, V., Perez, E. A., Saphner, T., Wolff, A. C. & Sledge-Jr, G. W. (2008). Weekly paclitaxel in the adjuvant treatment of breast cancer. *New England Journal of Medicine*, **358**(16), 1663–1671.

Spiegelhalter, D. J., Thomas, A., Best, N. G. & Gilks, W. R. (1999). *WinBugs: Bayesian inference using Gibbs sampling*. MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**(4), 583–639.

Sy, J. P. & Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, **56**(1), 227–336.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Thomas, A., O'Hara, B., Spiegelhalter, D., Best, N., Lunn, D. & Rice, K. (2007). Openbugs and its R / S-PLUS interface BRugs. http://mathstat.helsinki.fi/openbugs/.

Tucker, S., Thames, H. & Taylor, J. (1990). How well is the probability of tumor cure after fractionated irradiation described by Poisson statistics? *Radiation Research*, **124**(3), 273–282.

Wei, L. J., Lin, D. Y. & Weissefeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association: Theory and Methods*, **84**(408), 100–116.

Yakovlev, A. Y. & Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, Singapore.

Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis os permanent employment in Japan. *Journal of the American Statistical Association*, **87**(418), 284–292.

Yannaros, N. (1994). Weibull renewal process. *Annals of the Institute of Statistical Mathematics*, **46**(4), 641–648.

Yin, G. & Ibrahim, J. (2005). A general class of bayesian survival models with zero and nonzero cure fractions. *Biometrics*, **61**(2), 403–412.

Yu, B. (2008). A frailty mixture cure model with application to hospital readmission data. *Biometrical Journal*, **50**(3), 386–394.

Zaider, M., Zelefsky, M. J., Hanin, L. G., Tsodikov, A. D., Yakolev, A. Y. & Leibel, S. A. (2001). A survival model for fractionated radiotherapy with an application to prostate cancer. *Physics in Medicine and Biology*, **46**(10), 2745–2758.