

**MÉTODOS BAYESIANOS PARA ANÁLISE DE  
DADOS DE EVENTOS RECORRENTES  
CONSIDERANDO UMA CLASSE GERAL DE  
MODELOS COM FRAGILIDADE  
MULTIPLICATIVA**

**GRAZIELA DUTRA ROCHA GOUVÊA**

**2010**

**GRAZIELA DUTRA ROCHA GOUVÊA**

**MÉTODOS BAYESIANOS PARA ANÁLISE DE DADOS DE EVENTOS  
RECORRENTES CONSIDERANDO UMA CLASSE GERAL DE  
MODELOS COM FRAGILIDADE MULTIPLICATIVA**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de “Doutor”.

Orientador

Prof. Dr. João Domingos Scalon

LAVRAS  
MINAS GERAIS – BRASIL  
2010

**Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da  
Biblioteca Central da UFLA**

Gouvêa, Graziela Dutra Rocha.

Métodos bayesianos para análise de dados de eventos recorrentes considerando uma classe geral de modelos com fragilidade multiplicativa / Graziela Dutra Rocha Gouvêa. – Lavras : UFLA, 2010.

137 p. : il.

Tese (doutorado) – Universidade Federal de Lavras, 2010.

Orientador: João Domingos Scalon.

Bibliografia.

1. Inferência bayesiana. 2. Análise de sobrevivência. 3. Metrópolis-Hastings. 4. Hemodiálise. II. Universidade Federal de Lavras. II. Título.

CDD – 519.542

**GRAZIELA DUTRA ROCHA GOUVÊA**

**MÉTODOS BAYESIANOS PARA ANÁLISE DE DADOS DE EVENTOS  
RECORRENTES CONSIDERANDO UMA CLASSE GERAL DE  
MODELOS COM FRAGILIDADE MULTIPLICATIVA**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de “Doutor”.

APROVADA em 22 de fevereiro de 2010

Profª. Dra. Vera Lúcia Damasceno Tomazella	UFSCAR
Prof. Dr. José Carlos Fogo	UFSCAR
Prof. Dr. Joel Augusto Muniz	UFLA
Prof. Dr. Marcelo Ângelo Cirillo	UFLA

Prof. Dr. João Domingos Scalon  
UFLA  
(Orientador)

LAVRAS  
MINAS GERAIS – BRASIL

*Aos meus pais, Adimaldo e Maria José,  
Dedico*

*“Para estar junto não é preciso estar perto,  
e sim do lado de dentro”.*  
*(Leonardo da Vinci)*

*“Você não sabe o quanto eu caminhei.  
Pra chegar até aqui...  
Percorri milhas e milhas antes de dormir,  
Eu nem cochilei  
Os mais belos montes escalei,  
Nas noites escuras de frio chorei ei, ei....”*

*Trecho da música A Estrada  
Composição: Toni Garrido/Lazão/ Da Gama*

## AGRADECIMENTOS

Agradeço a Deus, à Nossa Senhora Aparecida e à Nossa Senhora das Graças, por terem me dado força, saúde, coragem e vontade para chegar até aqui.

Agradeço aos meus pais, pelo apoio, pelo amor incondicional e pela confiança em mim depositada. Aos meus irmãos, Sanderson, Maldon, Sandra e Adalgiza, que também estiveram ao meu lado, me incentivando.

Aos meus sobrinhos, Yann, Yaskara, Sanderson Filho e Laila, pelo sorriso, pela alegria e pelo carinho. Também aos meus cunhados Pedro, Daniela, Pollyane e Cristiano e à minha amiga Landinha, presentes em todos os momentos.

Agradeço ao Fernando que, eu posso dizer, é muito mais que um namorado. É amigo, companheiro e conselheiro. É muito bom compartilhar os momentos da minha vida com você. Obrigada por tudo.

Agradeço ao Sr. João Batista e a D. Benedita, por terem me recebido com tanto carinho. Também a D. Leuda e ao Sr. Wanderley, por terem me apoiado nos momentos em que precisei.

Gostaria de agradecer a todos os meus amigos, especialmente aqueles que dividiram comigo dúvidas, angústias e alegrias no decorrer da pós-graduação.

As amizades que fiz em Lavras foram tão fortes e verdadeiras que, tenho certeza, vão durar por toda a minha vida. Jamais me esquecerei dos momentos que passamos juntos. Agradeço a vocês: Verônica, Josi, Paty Mendes, Fabrícia, Popó, Paty Paranaíba, Luciane, Luciene, Rejane, Tatá, Natascha, Tiago, Renato, Anderson e Roberta. Também ao Fabrício, não só pela amizade, mas pelo apoio, principalmente no período da qualificação.

Em especial, gostaria de agradecer ao Ricardo, o meu amigo *Little*, que é, na verdade, *Big*, em se tratando de generosidade. Sem a sua ajuda computacional, tudo seria muito mais difícil.

Agradeço aos colegas da UFMT, que me incentivaram e torceram pelo meu sucesso.

Gostaria de agradecer ao professor Dr. Scalon, pelos conselhos e pela ajuda nos momentos em que precisei.

À professora Dra. Vera Tomazella, eu, sinceramente, nem sei como agradecer. Ela me adotou, me orientou e me fez crescer. Jamais poderei esquecer a sua bondade, o seu carisma e a sua competência. Muito obrigada mesmo.

Não poderia deixar de agradecer a todos os professores e funcionários do DEX/UFLA e também ao apoio financeiro do CNPq.

Gostaria de agradecer também à equipe médica do Centro de Terapia Renal Substitutiva da Santa Casa de Misericórdia, em Lavras, MG, em especial às funcionárias Cássia e Carla, pela colaboração, compreensão e incentivo à pesquisa científica.

Enfim, agradeço a todos que, de alguma forma, me ajudaram nessa escalada. Confesso que não foi fácil chegar até aqui, mas cheguei e estou muito, muito feliz.

## SUMÁRIO

	<b>Página</b>
RESUMO	i
ABSTRACT	iii
1 INTRODUÇÃO.....	01
1.1Objetivos.....	03
2 REFERENCIAL TEÓRICO.....	05
2.1 Análise de sobrevivência.....	05
2.1.1 Funções básicas em Análise de Sobrevivência.....	07
2.1.2 Construção da função de verossimilhança.....	09
2.2 Modelos paramétricos em análise de sobrevivência.....	10
2.2.1 Distribuição Exponencial.....	10
2.2.2 Distribuição Weibull.....	11
2.2.3 Distribuição Log-normal.....	13
2.2.4 Distribuição Gama.....	13
2.3Modelo de regressão de Cox.....	14
2.4 Processos de contagem.....	18
2.5 Eventos recorrentes.....	22
2.6 Modelos de fragilidade.....	29
2.6.1Distribuições para a variável de fragilidade.....	34
2.6.2Métodos de estimação.....	36
2.7 Classe geral de modelos com fragilidade multiplicativa.....	37
2.7.1Formulação Matemática.....	42
2.8 Inferência Bayesiana.....	46
2.8.1 Distribuições <i>a priori</i> .....	47
2.8.2 Distribuição <i>a posteriori</i> .....	50
2.8.3Intervalo de credibilidade.....	51

2.8.4 Métodos computacionais.....	52
2.8.5 Métodos de Monte Carlo via cadeias de Markov.....	53
2.8.5.1 Metropolis-Hastings.....	54
2.8.5.2 Amostrador de gibbs ( <i>Gibbs Sampling</i> ).....	55
2.8.6 Testes para avaliação da convergência.....	57
2.8.7 Critérios para seleção de modelos.....	59
3 MATERIAL E MÉTODOS.....	61
3.1 Modelo de Fragilidade Multiplicativo.....	62
3.2 Classe geral de modelos para eventos recorrentes com fragilidade multiplicativa.....	65
3.3 Análise bayesiana para o modelo de fragilidade multiplicativo.....	69
3.4 Análise bayesiana para a classe geral de modelos.....	71
3.5 Inferência a <i>posteriori</i> para as fragilidades.....	73
3.6 Aplicação.....	75
4 RESULTADOS E DISCUSSÃO.....	79
4.1 Classe geral de modelos para eventos recorrentes com fragilidade multiplicativa.....	79
4.2 Modelo de Fragilidade Multiplicativo.....	87
4.3 Modelo de Regressão de Cox.....	95
4.4 Comparação de modelos.....	98
5 CONCLUSÕES E TRABALHOS FUTUROS.....	99
REFERÊNCIAS BIBILOGRÁFICAS.....	101
ANEXOS.....	113

## RESUMO

GOUVÊA, Graziela Dutra Rocha. **Métodos bayesianos para análise de dados de eventos recorrentes considerando uma classe geral de modelos com fragilidade multiplicativa**. 2010. 137 p. Tese (Doutorado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras.

Em muitas situações há interesse em observar a ocorrência de um único tipo de falha, mais de uma vez, em cada indivíduo envolvido na análise. Nesse caso, é razoável supor a existência de associação entre tempos de ocorrência de eventos. Para considerar a existência de uma possível associação entre os tempos de sobrevivência, uma classe de modelos, denominada modelos de fragilidade, na qual um efeito aleatório é introduzido no modelo de Cox, vem sendo estudada. Peña e Hollander (2004) propuseram uma nova classe de modelos, mais geral e flexível, que incorpora simultaneamente os efeitos de covariáveis, o impacto da acumulação de tempos de ocorrências de eventos sobre o indivíduo, o efeito de variáveis latentes e o efeito das intervenções realizadas após cada ocorrência de evento. No que se refere à abordagem bayesiana, os modelos de fragilidade já foram explorados por vários autores. Por outro lado, a classe geral de modelos para eventos recorrentes ainda não foi estudada sob o enfoque bayesiano. Sendo assim, a proposta desta tese foi a utilização de métodos bayesianos para a classe geral de modelos com fragilidade para análise de dados de eventos recorrentes e também estudar, sob esse enfoque, o modelo de fragilidade multiplicativo e o modelo de Cox. Os modelos foram comparados considerando os métodos de seleção AIC (Critério de Informação de Akaike) e BIC (Critério de Informação Bayesiano). A metodologia proposta foi aplicada aos dados de Insuficiência Renal Crônica da Santa Casa de Misericórdia na cidade de Lavras, MG, os quais foram coletados exclusivamente para utilização nesta tese. Para esse conjunto de dados o efeito das covariáveis foi estudado e as fragilidades individuais para os pacientes em tratamento de Hemodiálise foram estimadas. Concluiu-se que a classe geral de modelos é importante porque considera além dos efeitos presentes nos outros modelos, o efeito da acumulação da recorrência dos eventos sobre os indivíduos. A aplicação ao conjunto de dados mostrou a importância dos termos de fragilidade e do parâmetro que mede

---

Comitê Orientador: João Domingos Scalon – UFLA (Orientador) e Vera Lúcia Damasceno Tomazella - UFSCAR

efeito da acumulação da recorrência dos eventos sobre os indivíduos. Foi mostrado que o modelo proposto por Peña e Hollander pode ser útil para analisar dados de tempos de recorrência de hospitalizações para pacientes em tratamento de Hemodiálise. Nesse caso, foi verificado que o efeito do acúmulo de recorrências de evento é um fator importante para predizer uma nova internação.

## ABSTRACT

GOUVÊA, Graziela Dutra Rocha. **Bayesian methods for data analysis of recurrent events considering a general class of models with multiplicative frailty**. 2010. 137 p. Thesis (Doctorate in Statistics and Agricultural Experimentation) – Federal University of Lavras, Lavras.

In some situations, we may observe that the event of interest occurs repeatedly in the same individual, such as when a patient diagnosed with cancer tends to relapse over time or when a person is repeatedly readmitted in a hospital. In this situation, it is more reasonable to suppose the existence of a relationship between times until the occurrence of events. In order to examine the association between survival times, a class of models has been studied. These are called frailty models and consider a random effect in the Cox model. Peña and Hollander (2004) proposed a new class of models, more general and flexible, that simultaneously incorporates the effect of covariates, the impact on the unit of accumulating event occurrences, the effect of latent or unobserved variables which, for each unit, endow correlation among the inter-event times, as well as the effect of performed interventions after each event occurrence. With respect to Bayesian approach, the frailty models have already been explored by many authors. By the other side, the general class of models for recurrent events has not been studied under Bayesian approach. Then, the aim of this thesis was to explore Bayesian methods for the general class of frailty models for recurrent event data analysis and also study, under this focus, the multiplicative frailty model and the Cox model. The models were compared using the selection methods AIC (Akaike information criterion) and BIC (Bayesian information criterion). The proposed methodology was applied to data of chronic renal failure from Santa Casa de Misericórdia in Lavras, MG, which was collected only for this study. The effect of covariates was studied and the individual frailties for the patients in treatment of hemodialysis were estimated. We conclude that the general class of models is important because consider, beside of effects in other models, the effect of cumulating times of recurrent events in the individuals. The application to the dataset showed the importance of both frailty terms and the parameter which measures the effect of cumulating times of recurrent events. It was showed that the model purposed by Peña and Hollander can be useful to analyse recurrent times data of hospitalization to patients in

---

Guidance Committee: João Domingos Scalon – UFLA (Master Professor) e Vera Lúcia Damasceno Tomazella - UFSCAR

treatment of hemodialysis. In this case, it was checked that the effect of cumulating recurrent events is an important feature to predict a new hospitalization.

## 1 INTRODUÇÃO

Em análise de sobrevivência considera-se, geralmente, que os tempos de ocorrência de eventos de indivíduos distintos são independentes. No entanto, existem diversas situações para as quais essa suposição não é adequada. Segundo Colossimo & Giolo (2006), situações em que é razoável supor a existência de associação entre tempos de sobrevivência caracterizam dados de sobrevivência multivariados.

A suposição de independência entre os tempos de falha não é válida, por exemplo, nos casos em que um indivíduo está sujeito a eventos recorrentes. Nesses casos, há interesse em observar a ocorrência de um único tipo de falha mais de uma vez em cada indivíduo envolvido na análise. Desse modo, mesmo tendo experimentado o evento, o indivíduo continua em observação, já que o mesmo evento pode ocorrer novamente. Exemplos de eventos recorrentes são episódios de gripe, infartos do miocárdio, diarreia, hospitalizações, entre outros. Nesses tipos de estudos, em que os dados de sobrevivência são correlacionados, o uso de métodos padrões para dados de tempo de falha podem levar a inferências inválidas. Sendo assim, incorporar a correlação pode levar a uma crescente eficiência na análise (Carvalho et al., 2005).

Para considerar a existência de uma possível associação entre os tempos de sobrevivência, uma classe de modelos, denominada modelos de fragilidade, inicialmente proposta por Clayton (1978), vem sendo estudada por diversos pesquisadores.

Os modelos de fragilidade são caracterizados pela introdução de um efeito aleatório, representado por uma variável aleatória não observável, no modelo de Cox (1972). Esse efeito aleatório, chamado fragilidade, representa as informações que não podem ou não foram observadas e tem como objetivo explicar a correlação entre os tempos de falha multivariados.

Existem vários modelos e métodos de análise de eventos recorrentes. No entanto, Peña & Hollander (2004) sugerem que ainda há a necessidade de uma classe de modelos mais geral e flexível que incorpore, simultaneamente, os efeitos de covariáveis, o impacto da acumulação de tempos de ocorrências de eventos sobre a unidade ou indivíduo, o efeito de variáveis latentes ou não observadas que, para cada indivíduo, denota a correlação entre os tempos interevento, bem como o efeito das intervenções realizadas após cada ocorrência de evento. Sendo assim, os autores propuseram uma nova classe de modelos com maior generalização do que a maioria dos atuais modelos baseados no modelo de Cox.

Geralmente, a estimação dos parâmetros dos modelos que envolvem a variável de fragilidade tem sido feita por meio do algoritmo EM (Esperança-Maximização), no entanto, alguns autores sugerem que a convergência em muitos casos é muito lenta.

A análise bayesiana é uma alternativa importante aos procedimentos clássicos de estimação que vem ganhando espaço crescente em aplicações nas mais diversas áreas. Segundo Tomazella (2003), a abordagem bayesiana é adequada quando ocorrem situações nas quais os dados avaliados são incompatíveis com modelos simples e outros modelos mais realísticos e, conseqüentemente, modelos mais complexos são necessários.

Do ponto de vista bayesiano, os modelos de fragilidade com distribuição gama já foram explorado por alguns autores. A grande dificuldade de se fazer inferência com modelos de fragilidade tem sido computacional e os métodos bayesianos minimizam esses problemas.

Por outro lado, a classe geral de modelos para eventos recorrentes com fragilidade multiplicativa, proposta por Peña & Hollander (2004), ainda não foi estudada sob o enfoque bayesiano. Como esse modelo é ainda mais complexo que o modelo de fragilidade usual, supõe-se que a metodologia bayesiana pode

ajudar a minimizar também as dificuldades computacionais impostas por tal modelo.

Os modelos de análise de sobrevivência são muito utilizados em pesquisas médicas e têm sido considerados bastante úteis no estudo de doenças crônicas.

Uma doença que constitui um grave problema médico e de saúde pública é a insuficiência renal (IR), caracterizada pela incapacidade dos rins em excretar substâncias tóxicas do organismo de forma adequada (Cardozo et al., 2006). As causas da insuficiência renal são muitas, algumas das quais acarretam uma diminuição rápida da função renal, muitas vezes, com valores inferiores a 1% ou 2% do índice normal (insuficiência renal aguda). Outras causas de IR acarretam a perda gradual e progressiva de grande parte dos néfrons funcionantes (Insuficiência Renal Crônica).

Dada a relevância e o aumento de casos de IR no Brasil (Sociedade Brasileira de Nefrologia - SBN, 2009), faz-se necessário o estudo de ferramentas estatísticas apropriadas que auxiliem na interpretação dos parâmetros dos modelos que podem ser aplicados na análise dos dados relacionados ao tratamento dessa doença. As técnicas de análise de sobrevivência são aqui consideradas, pois se adéquam cada vez mais aos dados que frequentemente são encontrados em vários tipos de estudos, especialmente os clínicos e observacionais. As técnicas bayesianas também são consideradas, pois, a cada dia, se mostram mais eficientes, especialmente no que diz respeito ao tratado computacional.

## **1.1 Objetivos**

A proposta, neste trabalho, é utilizar métodos bayesianos para a classe geral de modelos para análise de dados de eventos recorrentes com fragilidade multiplicativa (Peña & Hollander, 2004).

A metodologia proposta foi aplicada aos dados de insuficiência renal crônica da Santa Casa de Misericórdia, na cidade de Lavras, MG no intuito de determinar quais covariáveis estão relacionadas com um aumento e ou uma diminuição do tempo até a recorrência do evento de interesse e também estimar as fragilidades individuais para os pacientes em tratamento.

Também é de interesse comparar a classe geral de modelos para eventos recorrentes com fragilidade multiplicativa, proposta por Peña & Hollander (2004), e o modelo de fragilidade multiplicativo (Clayton, 1978) e o modelo de regressão de Cox (1972).

## 2 REFERENCIAL TEÓRICO

### 2.1 Análise de sobrevivência

Modelos de análise de sobrevivência objetivam estudar dados de experimentos em que, geralmente, a variável resposta é o tempo até a ocorrência de um evento de interesse (tempo de falha). Em muitas situações, no entanto, não é possível observar o tempo de falha para todos os indivíduos em estudo. Isso ocorre por vários motivos, dentre eles, a interrupção do acompanhamento de determinados indivíduos ou pelo fato de esses indivíduos não terem experimentado o evento até o término do estudo. Sendo assim, para determinados indivíduos, não é obtida uma resposta completa, ou seja, não se sabe se o evento ocorreu ou quando ele ocorreu. Segundo Kalbfleisch & Prentice (2002), essas são as principais características dos dados de sobrevivência: a presença de censura e a observação parcial da resposta.

Existem três tipos de censura: censura à direita, censura à esquerda e censura intervalar e três mecanismos de censura: censura tipo I, censura tipo II e censura aleatória. A ocorrência ou não de censura é geralmente representada por meio de uma variável indicadora de falha, que vale 1 se a observação falhou e 0 se foi censurada (Carvalho et al., 2005).

Segundo Colosimo & Giolo (2006), censura tipo I é aquela em que o estudo irá terminar após um período de tempo pré-estabelecido pelo pesquisador, enquanto a censura do tipo II é aquela em que o estudo termina após o evento de interesse ter ocorrido o número de vezes pré-estabelecido. A censura do tipo aleatória é aquela na qual a unidade em observação deixa de ser acompanhada sem que tenha experimentado o evento de interesse, por motivos alheios ao estudo.

A censura à direita é caracterizada pelo fato de o tempo registrado para a ocorrência de evento ser maior do que o tempo que determina o final do estudo,

enquanto para a censura à esquerda os tempos de ocorrência de eventos registrados são anteriores ao início do estudo. Já a censura intervalar é caracterizada pelo fato de não ser conhecido o tempo exato de ocorrência de um evento, sabendo-se somente que esse ocorreu num determinado intervalo de tempo (Gouvêa, 2006).

A presença de censuras torna inviável o uso de técnicas estatísticas convencionais para a análise dos dados. A aplicação de modelos de regressão na análise de sobrevivência também necessita de técnicas específicas para tratar as censuras. Assim, alguns métodos nessa área têm sido desenvolvidos e o interesse pelo assunto cresceu muito após a publicação do artigo de Cox (1972), no qual é apresentado o modelo de riscos proporcionais, que modela a função de risco na presença de covariáveis.

Além do tempo de sobrevivência e da variável indicadora de falha, os dados de sobrevivência podem conter um conjunto de variáveis observáveis que podem estar relacionadas com esses tempos. Essas variáveis são conhecidas por covariáveis ou variáveis explicativas. Quando os tempos de sobrevivência estão relacionados com as covariáveis diz-se que a população é heterogênea. Caso contrário, a população é dita homogênea (Bastos, 2003).

Geralmente, em análise de sobrevivência, considera-se que os tempos de ocorrência de eventos de indivíduos distintos são independentes. No entanto, existem diversas situações para as quais essa suposição não é adequada. Em estudos com animais, por exemplo, é razoável supor que exista correlação entre os tempos dos animais de uma mesma ninhada, enquanto em estudos com humanos pode haver correlação entre os tempos de ocorrência de eventos para os membros de uma mesma família, devido a influências genéticas ou a ambientes compartilhados. Esses tipos de dados são denominados tempos de falha multivariados ou dados multivariados de sobrevivência (Giolo, 2003).

Outro exemplo ocorre quando cada indivíduo em estudo está sujeito a múltiplos eventos do mesmo tipo (eventos recorrentes) tais como gripe, infartos do miocárdio, diarreia, entre outros. Eventos de tipos diferentes, tais como sequelas múltiplas em pacientes com doenças crônicas, são também exemplos nos quais a suposição de independência dos tempos pode não ser válida (Colossimo & Giolo, 2006).

### 2.1.1 Funções básicas em análise de sobrevivência

Seja  $T \geq 0$ , uma variável aleatória não negativa, contínua que representa o tempo de falha. Geralmente, essa variável é especificada, em análise de sobrevivência, pela sua função densidade de probabilidade, pela função de sobrevivência ou, ainda, pela função taxa de falha, também conhecida como função de risco ou, simplesmente, função risco.

A função de distribuição de  $T$  é definida como:

$$F(t) = P(T \leq t)$$

e a sua função densidade de probabilidade (*fdp*) é dada por:

$$f(t) = \frac{\partial}{\partial t} F(t)$$

A *fdp* pode ser representada em termos da probabilidade de ocorrência de um evento num intervalo de tempo infinitesimal de tamanho  $dt$ . Assim, tem-se:

$$f(t) = \lim_{dt \rightarrow 0} \frac{P(T \leq t + dt) - P(T \leq t)}{dt} \quad (2.1)$$

Note que a definição (2.1) é, de fato, a definição da derivada de  $F(t)$ .

A função de sobrevivência é definida como a probabilidade de uma unidade em observação não falhar até certo tempo  $t$ , pré-determinado. Em outras palavras, a probabilidade de uma unidade em observação sobreviver ao tempo  $t$ . Assim,

$$S(t) = P(T \geq t) = \int_t^{\infty} f(u) du \quad (2.2)$$

ou, ainda,

$$S(t) = 1 - F(t)$$

A função de risco é definida pelo limite da probabilidade de uma unidade observada falhar no intervalo de tempo  $[t, t + dt)$ , para  $dt$  tendendo a zero, dado que a unidade tenha sobrevivido até o tempo  $t$ . A função risco é dada pela expressão:

$$\lambda(t) = \lim_{dt \rightarrow 0^+} \frac{P(t < T \leq t + dt | T \geq t)}{dt} \quad (2.3)$$

Segundo Klein & Moeschberger (1997), a função de risco é bastante útil para se determinar a distribuição dos tempos de vida e para descrever o modo como as chances de ocorrência de um evento mudam com o tempo. Quanto às suas características, podem-se observar diferentes tipos de funções risco, de acordo com a forma do gráfico de  $\lambda(t)$ . A função risco pode ser constante, crescente, decrescente ou não monótona.

A função de risco pode ser definida, em termos de (2.1) e (2.2), por meio das expressões:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{\partial}{\partial t} \log S(t) \quad (2.4)$$

e

$$S(t) = \exp \left[ -\int_0^t \lambda(u) du \right] = \exp[-\Lambda(t)] \quad (2.5)$$

em que  $\Lambda(t) = \int_0^t \lambda(u) du$  é a função de risco acumulada, definida para algum  $t \geq 0$  e  $\int_0^{\infty} \lambda(u) du = \infty$ .

Utilizando (2.4) e (2.5), a *fdp* pode ser reescrita em termos da função risco, como segue:

$$f(t) = \lambda(t)S(t) = \lambda(t) \exp \left[ -\int_0^t \lambda(u) du \right] \quad (2.6)$$

### 2.1.2 Construção da função de verossimilhança

Para um conjunto de dados de sobrevivência  $D$  com  $n$  indivíduos, considere que o tempo de vida do  $i$ -ésimo indivíduo seja representado pela variável aleatória  $T_i$  e que, associado a esse indivíduo, existe um tempo de censura fixo  $C_i$ .

A representação dos dados sujeitos a censuras à direita do tipo I, para populações homogêneas, é dada por meio do par de variáveis aleatórias  $(t, \delta)$ , de modo que, para o  $i$ -ésimo indivíduo em observação, os dados de sobrevivência são dados por  $(t_i, \delta_i)$ ,  $i = 1, 2, \dots, n$ .

O que se observa para cada indivíduo é, então:

$$t = \min(T, C)$$

e

$$\delta_i = \begin{cases} 1 & \text{se } T \leq C \\ 0 & \text{se } T > C \end{cases}$$

Os tempos de sobrevivência  $t_1, t_2, \dots, t_n$  são assumidos independentes e identicamente distribuídos (i.i.d), com função densidade de probabilidade  $f(t|\theta)$ , em que  $\theta$  é o vetor de parâmetros (Lawless, 1982; Colosimo & Giolo, 2006).

A função de verossimilhança para os dados de sobrevivência sujeitos a censuras à direita do tipo I é dada por

$$L(\theta|D) = \prod_{i=1}^n f(t_i, \delta_i|\theta) = \prod_{i=1}^n [f(t_i|\theta)]^{\delta_i} [S(t_i|\theta)]^{1-\delta_i} \quad (2.7)$$

Por meio da função de verossimilhança (2.7), pode-se observar que a contribuição de cada observação não-censurada é a sua função de densidade e que a contribuição de cada observação censurada é a sua função de sobrevivência (Colosimo & Giolo, 2006).

Utilizando-se as relações básicas de sobrevivência, pode-se obter uma representação da função de verossimilhança expressa unicamente em termos da

função risco. Ou seja, a função risco especifica completamente a função de verossimilhança.

Assim, pode-se reescrever a expressão (2.7) como segue:

$$L(\boldsymbol{\theta}|\mathcal{D}) = \left\{ \prod_{i=1}^n [\lambda(t_i|\boldsymbol{\theta})]^{\delta_i} \right\} \exp \left\{ - \int_0^{t_i} \lambda(u|\boldsymbol{\theta}) du \right\} \quad (2.8)$$

Para populações heterogêneas, o conjunto  $\mathcal{D}$  é formado por:

$$\mathcal{D} = \{(t_i, \delta_i, x_i), i = 1, 2, \dots, n\}$$

em que  $t_i$  é o valor observado do tempo de sobrevivência,  $\delta_i$ , a função indicadora de censura e  $x_i$  é o vetor de covariáveis observadas para o  $i$ -ésimo indivíduo.

A função risco é dependente do vetor de parâmetros  $\boldsymbol{\theta}$  acrescido dos parâmetros referentes ao vetor de covariáveis. Sendo assim, a função de verossimilhança, dada em (2.8), pode ser reescrita como:

$$L(\boldsymbol{\theta}|\mathcal{D}) = \left\{ \prod_{i=1}^n [\lambda(t_i|x_i, \boldsymbol{\theta})]^{\delta_i} \right\} \exp \left\{ - \int_0^{t_i} \lambda(u|x_i, \boldsymbol{\theta}) du \right\} \quad (2.9)$$

## 2.2 Modelos paramétricos em análise de sobrevivência

Para a análise estatística de dados de sobrevivência podem ser utilizadas distribuições de probabilidade. Há vários modelos probabilísticos utilizados para descrever tempos de vida. Dentre eles destacam-se o modelo exponencial, o modelo Weibull e o modelo lognormal, além da distribuição gama, devido à adequação dos mesmos a diversas situações práticas.

### 2.2.1 Distribuição exponencial

Historicamente, segundo Carvalho et al. (2005), a distribuição exponencial foi a mais amplamente utilizada para modelar o tempo de vida. Segundo as autoras, ao final dos anos 1940, a distribuição exponencial era

utilizada em estudos de confiabilidade de sistemas eletrônicos e, posteriormente, em estudos de tempos de vida e remissão em doenças crônicas e infecciosas.

A função densidade de probabilidade para a variável aleatória tempo de falha  $T$ , que segue uma distribuição exponencial com parâmetro  $\alpha \geq 0$ , é dada por:

$$f(t) = \alpha \exp(-\alpha t) \quad (2.10)$$

Sua função de sobrevivência é dada pela expressão:

$$S(t) = \exp(-\alpha t) \quad (2.11)$$

Segundo Lawles (1982), como a distribuição exponencial é caracterizada pela falta de memória, a função de risco não depende do tempo. Nesse caso, a função de risco é expressa por

$$\lambda(t) = \alpha, \quad \text{para } t \geq 0. \quad (2.12)$$

### 2.2.2 Distribuição Weibull

A distribuição Weibull tem sido uma das mais utilizadas para modelar dados de sobrevivência nos últimos anos (Tomazella, 2003). A sua popularidade em aplicações práticas deve-se ao fato de tal distribuição apresentar uma grande variedade de formas, todas elas com uma propriedade básica: a sua função taxa de falha é monótona, ou seja, ela é crescente, decrescente ou constante no tempo.

Assumindo que o tempo de vida, dado pela variável aleatória  $T$ , apresenta uma distribuição Weibull, a sua função densidade de probabilidade é dada por:

$$f(t) = \gamma \alpha^\gamma t^{\gamma-1} \exp(-(\alpha t)^\gamma) \quad (2.13)$$

em que  $\gamma$ , o parâmetro de forma, e  $\alpha$ , o parâmetro de escala, são ambos positivos.

Sua função de sobrevivência é dada por:

$$S(t) = \exp(-(\alpha t)^\gamma) \quad (2.14)$$

Diferentemente do caso exponencial, a função risco varia com o tempo de acordo com a expressão:

$$\lambda(t) = \gamma \alpha^\gamma t^{\gamma-1} \quad (2.15)$$

Pode-se observar que, quando  $\gamma = 1$ , tem-se uma distribuição exponencial, logo, a distribuição exponencial é um caso particular da distribuição Weibull. Portanto, quando se tem  $\gamma = 1$ , a função risco é constante (como no modelo exponencial). Já no caso em que  $\gamma < 1$ , a função risco é decrescente e quando  $\gamma > 1$ , a função risco é crescente, como pode ser visto na FIGURA 1.

A função risco acumulado da distribuição Weibull é dada por:

$$\Lambda(t) = -\ln S(t) = (\alpha t)^\gamma$$

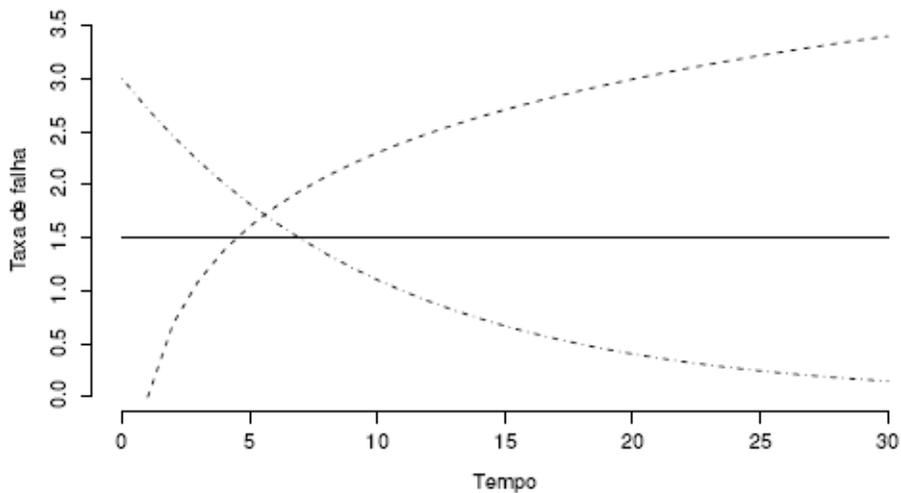


FIGURA 1 Funções de taxa de falha -- crescente, — constante, - · - constante.

### 2.2.3 Distribuição Log-normal

Assim como a distribuição Weibull, a distribuição Log-normal pode ser utilizada para caracterizar tempos de vida de produtos e indivíduos. Segundo Colosimo & Giolo (2006), essa distribuição também é utilizada para descrever situações clínicas, como, por exemplo, o tempo de vida de pacientes com leucemia.

Assumindo-se que a variável aleatória  $T$  segue uma distribuição Log-normal, isto é, que o logaritmo de  $T$  segue uma distribuição normal com parâmetros  $\mu$  e  $\sigma^2$ , tem-se:

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{1}{2} \left( \frac{\ln(t) - \mu}{\sigma} \right)^2 \right] \quad (2.16)$$

As funções de sobrevivência e de risco nesse caso envolvem integração incompleta da distribuição normal padrão (Carvalho et al., 2005). A expressão para a função de sobrevivência é dada por:

$$S(t) = 1 - \Phi \left( \frac{\ln(t) - \mu}{\sigma} \right) \quad (2.17)$$

em que  $\Phi$  é a função de distribuição acumulada da distribuição normal.

A função de risco pode, então, ser encontrada utilizando-se a expressão (2.4), uma vez que, no caso da distribuição Log-normal são conhecidas a função densidade de probabilidade e a função de sobrevivência.

### 2.2.4 Distribuição Gama

A distribuição Gama, que também inclui a exponencial como caso particular, foi utilizada por Brown & Flood (1947) para descrever o tempo de vida de copos de vidro circulando em uma cafeteria. Por se ajustar a uma variedade de fenômenos, segundo Carvalho et al. (2005), a distribuição Gama

tem sido utilizada para descrever o tempo de vida de produtos e, mais recentemente, também na área médica.

A função densidade da distribuição Gama é caracterizada por dois parâmetros,  $k$  e  $\alpha$ , em que  $k > 0$  é chamado parâmetro de forma e  $\alpha > 0$ , é chamado parâmetro de escala, e é expressa por:

$$f(t) = \frac{1}{\Gamma(k)\alpha^k} t^{k-1} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \quad t > 0 \quad (2.18)$$

em que  $\Gamma(k)$ , a função Gama é definida por  $\Gamma(k) = \int_0^{\infty} x^{k-1} \exp(-x) dx$ .

A função de sobrevivência dessa distribuição é dada por:

$$S(t) = \int_t^{\infty} \frac{1}{\Gamma(k)\alpha^k} u^{k-1} \exp\left\{-\left(\frac{u}{\alpha}\right)\right\} du \quad (2.19)$$

Segundo Lawless (1982), a função taxa de falha para a distribuição Gama apresenta um padrão crescente ou decrescente, convergindo para um valor constante quando  $t$  cresce de 0 ao infinito.

### 2.3 Modelo de regressão de Cox

Os primeiros modelos de regressão para análise de sobrevida foram desenvolvidos na década de 1960 (Harris & Albert, 1991) e eram totalmente paramétricos. Em 1972, Cox desenvolveu um modelo de regressão semiparamétrico, também conhecido como modelo de riscos proporcionais de Cox, modelo de regressão de Cox ou, simplesmente, modelo de Cox.

O modelo de regressão de Cox, segundo Colosimo & Giolo (2006), permite a análise de dados provenientes de estudos de tempo de vida em que a resposta é o tempo até a ocorrência de um evento de interesse, ajustado por covariáveis. Esse modelo é amplamente utilizado na análise de regressão de dados de sobrevivência e representa um grande avanço na análise de dados censurados. Nele, a distribuição do tempo de vida é descrita pela função risco.

Considere  $p$  covariáveis, de modo que  $X$  seja um vetor com os componentes  $X = (x_1, x_2, \dots, x_p)'$ . A expressão do modelo de regressão de Cox é dada por:

$$\lambda(t|X) = \lambda_0(t)g(X'\beta) \quad (2.20)$$

em que  $g$  é uma função não negativa que deve ser especificada, tal que  $g(0) = 1$  e  $\beta$  é o vetor de parâmetros associado às covariáveis. Este modelo é composto pelo produto de dois componentes de naturezas distintas, um não-paramétrico e outro paramétrico, o que o torna bastante flexível. O componente não-paramétrico  $\lambda_0(t)$  não é especificado e é uma função não negativa do tempo. Ele é usualmente denominado função de base ou basal, pois  $\lambda(t) = \lambda_0(t)$  quando  $X = 0$ . A parte paramétrica do modelo,  $g(X'\beta)$ , assume, geralmente, a forma  $\exp(X'\beta)$ .

Sendo assim, o modelo de riscos proporcionais de Cox pode ser reescrito da seguinte forma:

$$\lambda(t|X) = \lambda_0(t) \exp(X'\beta) \quad (2.21)$$

O termo  $\exp(X'\beta)$  no modelo de Cox é interpretado como um risco relativo, uma vez que apresenta a razão entre os riscos de ocorrência do evento de interesse para uma unidade com covariável  $X$  e uma unidade padrão.

A suposição básica para o uso do modelo de regressão de Cox é que as taxas de falha sejam proporcionais. Neste caso, para fazer a verificação da proporcionalidade, pode ser utilizado o gráfico do logaritmo da função risco acumulado versus o tempo de cada covariável. De acordo com Allison (1995), a situação extrema de violação para essa suposição é caracterizada quando as curvas se cruzam.

O modelo de regressão de Cox é caracterizado pelos coeficientes  $\beta$ 's que medem os efeitos das covariáveis sobre a função taxa de falha. Dessa maneira, é necessário um método de estimação para se fazer inferências no modelo. O método de máxima verossimilhança usual não pode ser utilizado

nesse caso, pois a presença do componente não paramétrico  $\lambda_0(\mathbf{t})$  na função de verossimilhança o torna inapropriado.

Sendo assim, Cox (1975), para a estimativa do vetor  $\beta$ , propôs o método de máxima verossimilhança parcial. Na verossimilhança parcial, considera-se apenas a informação das unidades sob risco no tempo  $\mathbf{t}$  por meio do vetor de covariáveis  $\mathbf{X}$ , pois a função risco de base  $\lambda_0(\mathbf{t})$  é eliminada devido à propriedade de proporcionalidade.

Considerando que, em uma amostra com  $n$  indivíduos, existam  $k \leq n$  falhas distintas nos tempos  $t_1 < t_2 \dots < t_k$ , a verossimilhança parcial, baseada no modelo de riscos proporcionais (2.21), é expressa por:

$$L(\beta) = \frac{\exp(x_i' \beta)}{\sum_{j \in R(t_i)} \exp(x_j' \beta)} \quad (2.22)$$

em que  $R(t_i)$  é o conjunto de unidades em risco no instante  $t_i$ , ou seja, é o conjunto de unidades em observação e não censuradas no instante imediatamente anterior a  $t_i$ .

Considerando os dados na forma  $(t_i, \delta_i, x_i)$ , a verossimilhança parcial dada em (2.22) pode ser escrita na forma:

$$L(\beta) = \prod_{i=1}^n \left( \frac{\exp(x_i' \beta)}{\sum_{j \in R(t_i)} \exp(x_j' \beta)} \right)^{\delta_i} \quad (2.23)$$

As estimativas de máxima verossimilhança dos parâmetros  $\beta$ 's são obtidas maximizando-se o logaritmo de (2.23) por meio de métodos numéricos. Segundo Fogo (2007), essa maximização, geralmente, é alcançada utilizando-se o método de Newton-Raphson.

O modelo de regressão de Cox pode ser tratado parametricamente. Nesse caso,  $\lambda_0(\mathbf{t})$  assume uma distribuição de probabilidade paramétrica. Tal modelo, como dito anteriormente, é amplamente utilizado para a análise de sobrevivência. No entanto, assume que os acontecimentos ocorrem de modo

independente, ou seja, que o tempo e a ocorrência de eventos não estão relacionados com a formação inicial e as subsequentes ocorrências. Em especial, duas características de processos de eventos recorrentes violam o pressuposto de independência, segundo Cook & Lawless (2007), a heterogeneidade entre indivíduos e a dependência de eventos.

A heterogeneidade é produzida porque alguns indivíduos têm maior (ou menor) taxa de ocorrência de evento do que outros, devido a efeitos desconhecidos ou não mensuráveis. Por exemplo, alguns pacientes têm uma susceptibilidade genética a desenvolver certas doenças, vivenciando a sua primeira, segunda, terceira, enésima recaída mais rapidamente do que o resto da população.

Qualquer correlação entre os eventos podem acarretar duas consequências importantes. Em primeiro lugar, as estimativas são ineficientes, levando a estimativas de erros padrão incorretas. Esses problemas podem levar a conclusões errôneas em relação à significância estatística dos efeitos dos tratamentos. Em segundo lugar, a violação do pressuposto de independência pode induzir a vieses nos efeitos estimados. Aalen (1988) mostra que a heterogeneidade não observada produz estimativas de efeitos de tratamentos enfraquecidas.

A dependência entre os tempos de falha é generalizada por mecanismos para os quais, em cada indivíduo, o tempo de uma falha influencia no risco de outras. Muitos modelos de sobrevivência com base no modelo de riscos proporcionais Cox têm sido propostos para lidar com esse tipo de dados (Therneau & Grambsch, 2000; Therneau & Hamilton, 1997; Hosmer & Lemeshow 1989; Kelly & Lim, 2000; Barai & Teoh, 1997).

Segundo Aalen et al. (2004), o modelo de riscos proporcionais de Cox foi, então, estendido e desenvolvido num contexto mais geral. Nesse contexto,

são dadas duas abordagens: modelos marginais e condicionais, dos quais o último inclui os modelos de fragilidade, que serão abordados mais adiante.

#### 2.4 Processos de contagem

A formulação por processos de contagem e teoria *martingale* desempenha papel fundamental na teoria moderna de análise de sobrevivência, como pode ser visto em Andersen et al. (1993).

O desenvolvimento de modelos estatísticos baseados em processo de contagem para analisar dados de eventos recorrentes foi originalmente introduzido por Aalen (1978). O autor combinou elementos de integração estocástica, teoria *martingale* para tempos contínuos e teoria de processos de contagem numa metodologia que permite o desenvolvimento de técnicas de inferência para dados de sobrevivência na presença de censuras. Essa formulação permite acomodar com facilidade tanto dados com censuras quanto situações mais complexas, tais como, covariáveis tempo-dependentes, eventos recorrentes e eventos múltiplos.

Um processo de contagem  $\{N(t), t \geq 0\}$  é definido como o número de eventos de um processo pontual em  $[0, t]$ , ou seja,  $N(t)$  é um processo estocástico que assume os valores 0, 1, 2, ..., registrando um salto do valor  $i-1$  ao valor  $i$ , quando o evento de interesse ocorre pela  $i$ -ésima vez,  $i = 0, 1, 2, \dots$ . Até que se registre a primeira ocorrência de evento,  $N(t)$  é igual a zero, quando, então, salta para o valor 1. O processo continua até uma nova ocorrência de evento e, assim, vão se registrando saltos sucessivos de tamanho 1 até o encerramento das observações.

Um processo de contagem pode ser entendido como um conjunto de variáveis aleatórias  $\{N(t), t \geq 0\}$ , em que:

- (i)  $N(0) = 0$ ;
- (ii)  $N(t) < \infty$ ;
- (iii)  $N(t)$  é contínuo à direita;
- (iv)  $N(t)$  é uma função do tipo “escada”, com saltos de tamanho 1.

Em análise de sobrevivência, o caráter temporal de um processo de contagem requer o conhecimento dos indivíduos em cada instante  $t \in [0, \infty]$ . Para dados censurados à direita, essa informação no tempo  $t$  inclui o conhecimento das censuras e das falhas anteriores a  $t$ . Em alguns estudos, essa informação pode incluir os valores de um conjunto fixo de covariáveis, como, por exemplo, idade e sexo, além de valores de covariáveis dependentes do tempo (Fogo, 2007).

Esse conhecimento acerca do que aconteceu aos indivíduos até o instante  $t$  é chamado de história ou filtragem do processo de contagem no instante  $t$  e é representado por  $\mathcal{F}_t$ . No caso de dados censurados à direita, a história no tempo  $t$  consiste do conhecimento dos pares  $(T_i, \delta_i)$  para os quais  $T_i \leq t$ , para aqueles que experimentaram o evento e para os quais  $T_i > t$ , caso contrário. A filtragem do processo estocástico até um instante imediatamente anterior a  $t$ , denotada por  $\mathcal{F}_{t^-}$ , representa os dados disponíveis, referentes aos conjuntos de tempos de falhas e censuras, até um tempo  $t^- < t$  (Andersen et al., 1993).

Para a representação da teoria de processos de contagem define-se, inicialmente,  $dN(t)$  como a mudança em  $N(t)$  em um curto intervalo de tempo  $[t, t + dt)$ ,

$$dN(t) = N\left[(t + dt)^-\right] - N(t^-),$$

em que  $dN(t)$  assume os seguintes valores:

$$dN(t) = \begin{cases} 1, & \text{se uma falha ocorreu em } t, \\ 0, & \text{se não ocorreu falha em } t. \end{cases}$$

Segundo Tomazella (2003), no acompanhamento de  $n$  indivíduos, ao longo do tempo, até a ocorrência de um evento de interesse, considera-se que cada um desses indivíduos dá origem a um processo de contagem. Sob essa ótica, a partir do artigo de Aalen (1980), no qual foi sugerido um modelo de risco aditivo para analisar dados de sobrevivência, o par de variáveis  $(T_i, \delta_i)$ , em que  $\delta_i$  é o indicador de censura, é substituído pelo par de funções  $[N_i(t), Y_i(t)]$ ,  $i = 0, 1, \dots, n$ , em que  $N_i(t)$  indica o número de eventos observados para o indivíduo  $i$  no intervalo  $[0, t]$  e  $Y_i(t)$  é o processo indicador de risco, de modo que  $Y_i(t) = 1$ , se o indivíduo  $i$  encontra-se em risco no tempo  $t$  e  $Y_i(t) = 0$ , se o indivíduo  $i$  encontra-se fora de risco no tempo  $t$ . Os indivíduos que se encontram fora de risco são, normalmente, aqueles que já experimentaram o evento ou foram censurados.

Na FIGURA 2 encontra-se a representação gráfica para processos individuais. No gráfico da esquerda, pode-se observar a ocorrência de uma falha e, no gráfico à direita, o momento de uma censura. Observe que, em ambos os casos, os indivíduos saem de risco, no entanto, apenas no primeiro caso o evento de interesse é observado.

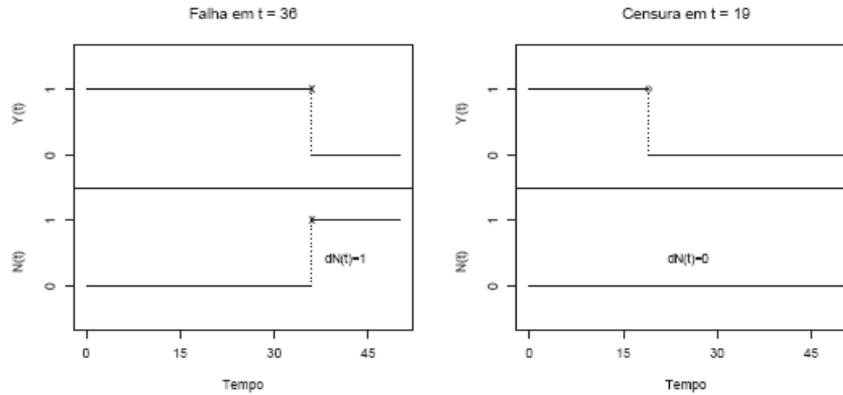


FIGURA 2 Representação gráfica individual, de falha (x) e censura (o), na forma de processos de contagem.

Os processos de contagem individuais podem ser convenientemente representados por um processo de contagem multivariado  $N(t) = [N_1(t), N_2(t), \dots, N_n(t)]$ . Então, o processo de contagem que corresponde ao total de eventos ocorridos até o tempo  $t$  é definido por  $N(t) = \sum_{i=1}^n N_i(t)$  e  $Y(t) = \sum_{i=1}^n Y_i(t)$  é o número de indivíduos em risco em  $t$ .

A estrutura probabilística de um processo de contagem pode ser descrita pelo processo de intensidade multivariado  $\lambda(t) = (\lambda_1(t), \lambda_2(t), \dots, \lambda_n(t))'$ , em que  $\lambda_i(t)$  é dado por:

$$\lambda_i(t) = \lim_{dt \rightarrow 0} \frac{P(N(t+dt) \geq t | \mathcal{F}_t^-)}{dt} \quad (2.24)$$

É importante comparar a função de risco (2.3) com a função intensidade (2.24). A função de risco é a probabilidade de que o evento de interesse ocorra apenas uma vez no intervalo de tempo  $[t, t+dt)$ , condicionada à sobrevivência no início do intervalo, dividida pelo tamanho do intervalo. A função intensidade é a probabilidade incondicional de ocorrência do evento (não necessariamente o

primeiro) no intervalo  $[t, t + dt)$ , dividida pelo tamanho do intervalo. Assim, apesar de diferentes, as funções de risco e intensidade possuem a mesma interpretação (Fogo, 2007). Vale ressaltar que a função risco será igual à função intensidade quando ocorre apenas uma falha no intervalo  $[t, t + dt)$ .

## 2.5 Eventos recorrentes

Segundo Cook & Lawless (2007), um evento é dito terminal quando a observação se encerra no exato momento de sua ocorrência. Eventos não terminais, por sua vez, ocorrem quando a observação não se encerra com a ocorrência do evento de interesse, podendo acontecer por diversas vezes durante o período de observação e são conhecidos como eventos recorrentes. Nesse caso, o indivíduo continua em observação mesmo depois da ocorrência de cada evento.

Os  $j$  tempos de observação de ocorrências de eventos ( $j = 1, 2, \dots, m_i$ ) para  $n$  indivíduos em observação, durante um período de tempo pré-estabelecido pelo pesquisador, estão representados na TABELA 1.

TABELA 1 Representação das  $m_i$  ocorrências de eventos para  $n$  indivíduos.

Indivíduo ( $i$ )	Tempos de observação de ocorrência de eventos ( $j$ )					
	1	2	3	4	...	$m_i$
1	$t_{11}$	$t_{12}$	$t_{13}$	$t_{14}$	...	$t_{1m_1}$
2	$t_{21}$	$t_{22}$	$t_{23}$	$t_{24}$	...	$t_{2m_2}$
3	$t_{31}$	$t_{32}$	$t_{33}$	$t_{34}$	...	$t_{3m_3}$
4	$t_{41}$	$t_{42}$	$t_{43}$	$t_{44}$	...	$t_{4m_4}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n$	$t_{n1}$	$t_{n2}$	$t_{n3}$	$t_{n4}$	...	$t_{nm_n}$

Em ciência e tecnologia, muitas vezes, o interesse reside em estudar processos que geram eventos repetidamente ao longo do tempo. Esses processos são conhecidos como processos de eventos recorrentes.

Segundo Cook & Lawless (2007), esses tipos de processos surgem frequentemente em estudos médicos, em que muitos indivíduos podem experimentar eventos clínicos repetidamente ao longo de um período de observação. Exemplos incluem a ocorrência de asma em estudos respiratórios, epilepsia em estudos de neurologia, osteoporose e fraturas, entre outros.

Segundo Tomazella (2003), eventos recorrentes são muito comuns em estudos de confiabilidade quando se observa a ocorrência de falhas de um equipamento ou máquina, durante um estudo de sua vida útil. Na pesquisa médica, podem-se encontrar dados de eventos recorrentes quando se estuda a reincidência de uma doença, como, por exemplo, gripes, tumores, pneumonias e ataques epiléticos, dentre outros.

O interesse em estudos que envolvem eventos recorrentes é caracterizar o processo de ocorrência de eventos para os indivíduos e também comparar tratamentos com base no tempo até a ocorrência de cada evento distinto. A idéia é explicar a natureza da variação entre indivíduos, covariáveis e outros fatores que não podem ser medidos.

Uma característica importante dos dados de eventos recorrentes, de acordo com Carvalho et al. (2005), é a dependência temporal ou estocástica, permitindo a modelagem desses eventos por meio de métodos de sobrevivência em processos de contagem.

Para analisar corretamente a ocorrência de eventos, é necessário dispor de uma escala de tempo adequada. Segundo Duchateau et al. (2003), para modelar a evolução temporal da taxa de falha em dados de eventos recorrentes, é crucial começar escolhendo a escala temporal que será utilizada. Dependendo da escala temporal selecionada, a interpretação da evolução temporal será totalmente diferente, tanto em modelos paramétricos quanto em modelos de fragilidade semiparamétricos.

A escala temporal também envolve a escolha de origem e isso requer alguns cuidados quando vários indivíduos se encontram em estudo. O desejável é que o pesquisador utilize uma origem que é coerente para todos os indivíduos e facilite a análise e interpretação dos dados. Em um estudo relacionado a pacientes com determinada doença, o tempo de origem poderia ser tomado como sendo o de início da doença. É uma escolha natural, embora esse tempo seja, muitas vezes, difícil de determinar precisamente.

Segundo Cook & Lawless (2007), em estudos clínicos aleatorizados, por exemplo, é habitual considerar o tempo de início do tratamento como o tempo origem, o que é bastante razoável, pois, em geral, o interesse é justamente fazer comparações entre tratamentos. Uma vez escolhida uma escala temporal, é necessário decidir se é mais adequado desenvolver modelos baseados em tempo total ou tempo entre eventos (*gap time*). Embora isso possa ser visto como uma decisão na especificação do modelo, afeta a análise e interpretação dos resultados.

Matuda (2005) afirma que, em muitos estudos, os dados são coletados durante algum período de tempo no calendário e os diferentes indivíduos não são necessariamente observados nos mesmos intervalos ao longo de seu processo de evento.

Existe vasta literatura sobre eventos recorrentes, como pode ser visto, por exemplo, em Wei et al. (1989), Pepe & Cai (1984), Lawless & Nadeaus (1995), Duchateau et al. (2003), Hoover (1996), Lee & Wang (2003), Cook & Lawless (2007).

Ghosh & Lin (2002) propuseram um modelo de regressão semiparamétrico que especifica o efeito multiplicativo das covariáveis sobre a função marginal. Para isso, concentraram-se na média marginal do número de eventos recorrentes ao longo do tempo. A função incorpora o fato de que um indivíduo que morre não pode experimentar mais eventos recorrentes.

Box-Steffensmeier & De Boef (2002) realizaram uma análise Monte Carlo sob uma variedade de condições experimentais para determinar quais modelos são os mais adequados para os dados de eventos recorrentes. Um estudo de simulação foi realizado com o objetivo de determinar, entre os modelos de variância robusta e de fragilidade, qual o que melhor explica a dependência de eventos recorrentes. Os resultados apontaram que os dois modelos são indicados para análise desse tipo de dados. As autoras reiteram que a escolha correta do modelo que melhor representa os dados, tais como a existência de eventos repetidos, pode fornecer grandes passos para evitar conclusões enganosas.

Segundo Jackson et al. (2002), as infecções respiratórias agudas, incluindo a gripe, têm sido sugeridas como possíveis precipitantes de eventos cardíacos agudos. Para avaliar os efeitos da vacinação contra eventos coronarianos recorrentes, estes autores realizaram um estudo de base populacional em uma coorte de 1.378 inscritos no Grupo Saúde Cooperativa que sobreviveram a um primeiro enfarte do miocárdio, de 1992 até 1996. O modelo de riscos proporcionais de Cox, com termo de fragilidade, foi utilizado para avaliar a associação da recepção de cada ano da vacina contra o risco subsequente de recorrentes infartos do miocárdio e morte por doença cardiovascular aterosclerótica.

Liu et al. (2004) propuseram um modelo semiparamétrico para a função intensidade de eventos recorrentes e um evento terminal por meio de um modelo de fragilidade compartilhada Gama. Os autores estenderam o modelo de fragilidade com censura informativa de Huang & Wolfe (2002) e aplicaram o método proposto em dados de hospitalização de pacientes candidatos a transplante de rins.

Segundo Mahé & Chevret (1999), as estimativas dos coeficientes de regressão obtidas por meio dos modelos de fragilidade somente são

interpretáveis condicionalmente à fragilidade não observada. Inversamente, modelos de riscos marginais não permitem medir o grau de dependência entre os tempos de falha no mesmo indivíduo, o que pode ser uma informação muito importante em estudos médicos. Os autores combinaram essas duas aproximações e afirmaram que esse seria o único caminho para estimar, simultaneamente, parâmetros de regressão marginais que representem o efeito médio das covariáveis e a dependência intra-unidades.

Chiang et al. (2005) consideraram o método dos momentos e o método de mínimos quadrados para estimar a função taxa de falha a partir de dados de eventos recorrentes. Com uma hipótese de censura independente no processo de eventos recorrentes, foram estudadas as propriedades estatísticas dos estimadores e propostos procedimentos *bootstrap* para a seleção da janela (núcleo estimadores) e para a aproximação de intervalos de confiança na estimativa da função taxa de ocorrência. Os autores declararam que, na execução dos procedimentos *bootstrap*, o método dos momentos é computacionalmente mais eficiente que o método de mínimos quadrados. O desempenho dos procedimentos propostos foi estudado por meio de simulações Monte Carlo e aplicado a um conjunto de dados epidemiológicos sobre usuários de drogas injetáveis.

Fredette & Lawless (2007) apresentaram métodos para prever eventos recorrentes que ocorrem em indivíduos ou em algumas unidades populacionais. Os eventos foram modelados utilizando o flexível processo de Poisson não homogêneo e uma possível heterogeneidade entre os indivíduos foi modelada utilizando efeitos aleatórios.

Em seu artigo, Huang & Liu (2007) consideraram uma das mais complicadas situações, aquela em que todos os eventos recorrentes estão sujeitos à censura por uma correlação com o evento terminal. Os autores avaliaram a

relação entre reincidência da doença e sobrevivência prolongada, utilizando um modelo baseado em *gap times*.

Segundo Ye et al. (2007), eventos recorrentes e eventos terminais estão fortemente correlacionados e, por esse motivo, os autores propuseram um modelo semiparamétrico, ao qual é incorporado um termo de fragilidade para modelar, conjuntamente, dados de eventos recorrentes e terminais. O modelo é semelhante ao modelo de fragilidade não paramétrico proposto por Huang & Wang (2004) e, em relação à análise, houve um notável aumento de eficiência. Embora um pressuposto paramétrico para a fragilidade seja feito no modelo proposto por Ye et al. (2007), os estudos de simulação sugerem que o modelo é robusto a desvios de suposição, pelo menos nos casos considerados.

Liu et al. (2004) utilizaram a estimativa de máxima verossimilhança em seu modelo de fragilidade, assumindo que os processos de eventos recorrentes seguem um processo de Poisson não homogêneo condicionado à fragilidade. Nesse estudo, um algoritmo EM/Monte Carlo/Metropolis-Hastings no passo-E foi adaptado para obter o estimador de máxima verossimilhança. No entanto, o algoritmo esperança-maximização, ou EM, tem convergência lenta e o método é muito mais intenso computacionalmente do que o método proposto por Ye et al. (2007).

De acordo com Cook & Lawless (2007), análises de eventos recorrentes podem ter vários objetivos, o que leva à formulação de diferentes métodos e modelos para análise dos dados. Nos experimentos em que os doentes são distribuídos aleatoriamente para tratamentos mediante a sua entrada no estudo e seguidos prospectivamente, por exemplo, comparações simples entre dois ou mais grupos de tratamento são frequentemente de interesse, baseadas em características marginais. Análises baseadas em contagem de eventos são muito úteis nesse contexto, pois proporcionam uma base para comparações de

tratamentos simples que exploram a aleatorização e facilitam inferências causais quanto aos efeitos dos tratamentos.

A modelagem de eventos recorrentes pode ser abordada de várias formas, as quais são descritas em livros sobre processos estocásticos. Para efeitos de ambas, modelagem e análise estatística, os conceitos de função intensidade e processos de contagem são especialmente úteis (Cook & Lawless, 2007, Lee & Wang (2003), Tomazella (2003), González (2006), entre outros.

Prentice et al. (1981) propuseram dois modelos para eventos recorrentes. Os dois modelos podem ser considerados como extensões do modelo de riscos proporcionais de Cox estratificado, sendo os estratos definidos pelo número e os tempos de recorrência de eventos. No primeiro modelo, o tempo de seguimento começa no início do estudo (verdadeiro tempo zero), enquanto, no segundo modelo, o tempo de seguimento começa no tempo de falha ou no tempo que precede imediatamente o evento.

O modelo proposto por Anderson & Gill (1982) assume que todos os eventos são do mesmo tipo e independentes. O conjunto de risco estabelecido na verossimilhança é totalmente diferente dos modelos propostos por Prentice et al. (1981). O conjunto de risco de um indivíduo no tempo de evento deverá conter todos os indivíduos que ainda estão sob observação, independente do número de eventos que tenham experimentado antes desse tempo.

Utilizando uma aproximação marginal, Wei et al. (1989) propuseram um modelo para a análise de falhas recorrentes. As falhas podem ser recorrências de um mesmo tipo de evento ou de eventos de diferentes naturezas, dependendo de como a estratificação é definida. Se os estratos são definidos pelos tempos de falhas repetidas do mesmo tipo de evento, de modo similar aos estratos definidos nos modelos propostos por Prentice et al. (1981), pode-se utilizá-lo para analisar falhas repetidas desse evento. Esse método modela a distribuição marginal de cada tempo de falha e não a estrutura de dependência sobre os distintos tempos

de falha aos quais os indivíduos foram submetidos. Nesse caso, cada recorrência é modelada como um estrato diferente.

## **2.6 Modelos de fragilidade**

Para considerar a existência de uma possível associação entre os tempos de sobrevivência, uma classe de modelos, denominada modelos de fragilidade, vem sendo estudada por diversos pesquisadores, dentre eles Oakes (1982), Hougaard (1986a; 1986b), Hougaard et al. (1992) e Hougaard (2000).

Os modelos de fragilidade são caracterizados pela introdução, no modelo de Cox (1972), de uma variável aleatória não observável, que representa as informações que não podem ou não foram observadas, como, por exemplo, fatores ambientais e genéticos, ou informações que, por algum motivo, não foram consideradas no estudo (Colosimo & Giolo, 2006).

Essa variável é introduzida no modelo quando a suposição de independência entre os tempos de ocorrência de eventos não é válida, com o objetivo de controlar a heterogeneidade não observável dos indivíduos em estudo, inclusive a dependência dos indivíduos que partilham os mesmos fatores de risco.

Modelos de fragilidade assumem que a distribuição dos efeitos individuais pode ser conhecida ou, pelo menos, aproximada, como sugerem Clayton (1978), Oakes (1982), Clayton & Cuzick (1985), Hougaard (1986a; 1986b), Andersen et al. (1993) e Hougaard (2000).

O primeiro trabalho nessa área foi atribuído a Clayton (1978), que propôs um estudo epidemiológico da incidência de doenças crônicas e analisou a tendência familiar na incidência dessas doenças. O termo fragilidade, indicando medida de associação, foi introduzido na análise de sobrevivência por Vaupel et al. (1979) e, desde então, vários autores consideram o modelo de fragilidade, tais

como Andersen et al. (1993), Ripatti & Palmgren (2000), Jensen et al. (2004) e Stefanescu & Turnbull (2006).

Hougaard (1986a) propôs uma classe de distribuições contínuas para o tempo de vida multivariado e estudou a dependência entre os indivíduos de um mesmo grupo. O autor afirmou que essa dependência pode ser interpretada como uma covariável não observada, comum aos indivíduos em estudo.

McGilchrist & Aisbett (1991) ajustaram um modelo de fragilidade para medidas repetidas e, para isso, utilizaram um conjunto de dados de pacientes de diálise. Nesse caso, foi estudado o tempo de reincidência de infecção após a inserção do cateter em pacientes que utilizam equipamento portátil de diálise. McGilchrist (1993) utilizou o mesmo conjunto de dados e propôs uma extensão do método utilizado por McGilchrist & Aisbett (1991), utilizando uma aproximação do melhor preditor linear não viesado como um passo inicial na computação da estimativa de máxima verossimilhança e da estimativa de máxima verossimilhança residual.

Kooperberg (1998) estudou uma coorte retrospectiva de gêmeos australianos para analisar a correlação entre gêmeos monozigóticos e gêmeos dizigóticos para várias doenças. Segundo o autor, o estudo se justifica pelo fato de gêmeos monozigóticos compartilharem ambiente na mesma placenta quando no útero da mãe, enquanto gêmeos dizigóticos são gerados em placentas separadas.

Parner (2001) propôs, como uma alternativa aos modelos aditivo e multiplicativo de fragilidade para dados multivariados de sobrevivência, uma inferência baseada numa função de verossimilhança composta, que requer somente os modelos das distribuições marginais dos pares de tempos de falhas. O autor fez um estudo envolvendo filhos e pais adotivos. O estudo foi constituído de 965 famílias. A ideia básica neste estudo de adoção foi que a semelhança, com respeito ao tempo de sobrevivência, entre as crianças adotivas

e pais biológicos, deve-se a fatores genéticos e a semelhança entre crianças adotivas e pais adotivos deve-se a fatores ambientais.

Jones & Rocke (2002) propuseram uma análise de sobrevivência multivariada envolvendo dados que apresentavam censura dupla. Os autores utilizaram o modelo marginal e o modelo de fragilidade para avaliar o tratamento contra câncer para fibroplasia óssea, uma doença muito rara, e ambos os modelos sugeriram que o tratamento foi efetivo. Kim (2006) investigou as diferentes situações em que a independência na ocorrência de eventos pode não ser verdadeira, focando a análise com censura dupla num estudo que envolvia hemofílicos contaminados pelo vírus da AIDS. Nesse trabalho, foi aplicado o modelo de fragilidade para descrever o efeito das covariáveis e o algoritmo EM foi desenvolvido para a estimação dos parâmetros em um estudo de simulação.

Clayton (1991) foi o primeiro a introduzir uma abordagem bayesiana nos modelos de fragilidade. Na abordagem bayesiana, a distribuição de fragilidade mais utilizada é a Gama, devido à sua flexibilidade e conveniência algébrica. Entretanto, outras distribuições de fragilidade têm sido pesquisadas, como, por exemplo, a família de distribuições estáveis positivas (Qiou et al., 1999).

Sinha (1993) apresenta uma análise bayesiana para dados de tempos de eventos múltiplos, baseada no modelo de intensidade proporcional. Sinha & Dey (1997) apresentam uma revisão completa dos modelos multiplicativos sob o enfoque bayesiano. Silva (2001) apresentou uma análise bayesiana para modelos de sobrevivência com fragilidade, sobretudo modelos de sobrevivência com estrutura aditiva para a função risco. Num estudo de câncer de pulmão, Dunson & Chen (2004) propuseram *prioris*, algoritmos computacionais e procedimentos para inferência bayesiana nos modelos de fragilidade compartilhada.

Os modelos de fragilidade podem ser utilizados tanto em estudos de sobrevivência univariados, quanto multivariados.

Num contexto univariado, cada indivíduo tem sua própria fragilidade e, nesse caso, pode-se pensar nesse indivíduo como sendo o representante de um grupo de tamanho 1. No entanto, se os indivíduos não apresentam tempos de sobrevivência multivariados, é razoável supor que não haja necessidade de considerar um componente de fragilidade para cada indivíduo. Para justificar a presença da fragilidade nesse caso, Colossimo & Giolo (2006) argumentam que, em estudos médicos, não importa quantas covariáveis sejam medidas, dois indivíduos com exatamente os mesmos valores das covariáveis não são esperados experimentar qualquer resposta médica exatamente no mesmo tempo. Esse comportamento pode ser explicado por diferenças genéticas, biológicas ou psicológicas que não podem ser mensuradas.

Diferentes abordagens têm sido propostas na literatura para análise de dados de sobrevivência multivariados, dentre os quais se destacam a abordagem condicional e a marginal. Os modelos de fragilidade, segundo Carvalho et al. (2005), classificam-se na abordagem condicional, uma vez que assumem que os tempos que apresentam uma possível associação são independentes condicionalmente às variáveis de fragilidade.

Para considerar a heterogeneidade dos indivíduos, que afeta de maneira significativa os tempos de sobrevivência observados, um efeito aleatório (fragilidade) é introduzido no modelo de Cox (2.21). Esse efeito aleatório, considerado em geral, não-negativo, é incorporado na função de risco como um fator multiplicativo. Segundo Giolo (2003), o fato de esse efeito atuar de forma multiplicativa na função de risco é, em princípio, arbitrário, mas tem sido utilizado na maioria dos trabalhos que envolvem dados de sobrevivência.

A interpretação do efeito aleatório é simples, no sentido de que quanto maior for seu valor, maior o risco de ocorrer uma falha, ou seja, mais frágeis os indivíduos de um grupo estarão para falhar.

Suponha que  $n$  indivíduos sejam observados, com  $j = 1, 2, \dots, m_i$  tempos de falha correlacionados para o  $i$ -ésimo indivíduo ( $i = 1, 2, \dots, n$ ). Sejam  $T_{ij}$  e  $C_{ij}$  os tempos de falha e de censura correspondente à  $j$ -ésima falha no  $i$ -ésimo indivíduo. Sendo assim, os dados observados consistem de  $\{(W_{ij}, Y_{ij}, X_{ij}) : i = 1, 2, \dots, n \text{ e } j = 1, 2, \dots, m_i\}$ , em que  $W_{ij} = \min(T_{ij}, C_{ij})$  são tempos de falha observados,  $Y_{ij} = I(T_{ij} \leq C_{ij})$ , isto é,  $Y_{ij} = 1$  se  $T_{ij} \leq C_{ij}$  e 0, caso contrário e, ainda,  $X_{ij}$  denotam os valores das covariáveis para a  $j$ -ésima observação para o  $i$ -ésimo indivíduo. Para garantir a identificabilidade, é assumido que o vetor tempo de falha  $T_i = (T_{i1}, \dots, T_{im_i})'$  é independente do vetor de tempos de censura  $C_i = (C_{i1}, \dots, C_{im_i})'$  condicional ao vetor de covariáveis  $X_{ij}$ .

O modelo de fragilidade estudado por Clayton & Cuzick (1985) formula a função risco da  $j$ -ésima falha no  $i$ -ésimo indivíduo como segue:

$$\lambda_{ij}(\tau | Z_i, X_{ij}) = Z_i \lambda_0(\tau) \exp(X_{ij}' \beta) \quad (2.25)$$

em que  $\lambda_0(\tau)$  é uma função risco de base,  $\beta$  é o vetor de coeficientes de regressão desconhecidos de dimensão  $p \times 1$ , e  $Z_i$  são variáveis aleatórias positivas, independentes e identicamente distribuídas, com distribuição conhecida  $g(z, \eta)$ , as quais são denominadas fragilidade.

Considerando as variáveis de fragilidade  $Z_i$ , a função de verossimilhança completa para o modelo (2.25) é dada por:

$$L = \prod_{i=1}^n g(z, \eta) \prod_{j=1}^{m_i} \{Z_i \lambda_0(T_{ij}) \exp(X_{ij}' \beta)\}^{Y_{ij}} \times \exp\{-\Lambda_0(T_{ij}) \exp(X_{ij}' \beta) Z_i\} \quad (2.26)$$

em que  $\Lambda_0(\cdot)$  é a função risco de base acumulada (Su & Fan, 2004).

Além da forma multiplicativa, as fragilidades podem ser introduzidas também de forma aditiva na função risco, como pode ser visto em Petersen

(1998), podendo, dessa maneira, responder a diferentes formas de avaliar a influência da heterogeneidade entre as unidades na função de risco ou da intensidade dos processos de contagem. Nesses modelos, os componentes de fragilidade são combinados aditivamente na função risco e, então, atuam multiplicativamente nas taxas de risco dos indivíduos.

Petersen (1998) sugeriu que uma abordagem geral para os tempos de falha multivariados deve levar em consideração variações devido a características não observadas compartilhadas e não compartilhadas. Sendo assim, o autor considerou modelos em que a associação entre os respectivos tempos de falha dos indivíduos de cada família deve-se, em parte, a relações familiares (fatores genéticos compartilhados) e, em parte, ao ambiente que esses indivíduos compartilham (fatores ambientais).

Tomazella (2003) realizou um estudo dos modelos de fragilidade, tanto multiplicativos quanto aditivos, em modelos de regressão para processos de Poisson, enquanto em Silva (2001), pode-se encontrar uma abordagem bayesiana pra esses modelos.

### **2.6.1 Distribuições para a variável de fragilidade**

Várias distribuições de probabilidade para a variável de fragilidade têm sido propostas na literatura, dentre elas a distribuição Gama, a Log-normal, a Gaussiana inversa e a Weibull.

Clayton (1978) e Oakes (1982) consideram o primeiro modelo de fragilidade para dados de sobrevivência multivariados, utilizando a distribuição Gama para a fragilidade. Whittmore & Lee (1991) estudaram o modelo com fragilidade Gama inversa.

Em trabalhos aplicados, a distribuição paramétrica mais assumida para fragilidades é a distribuição Gama (Vaupel et al., 1979; Clayton & Cuzick, 1985; Klein, 1992; Andersen et al., 1993). Isso se deve, principalmente, ao fato

de tal distribuição ser não negativa, extremamente flexível e algebricamente conveniente. No entanto, outras distribuições também têm sido utilizadas.

Hougaard (1984) sugeriu a distribuição Gaussiana inversa como alternativa à distribuição Gama para a variável de fragilidade, já que essa distribuição torna a população mais homogênea com o passar do tempo, o que é compatível com a ideia de que o evento de interesse ocorre primeiramente nas unidades mais frágeis. Hougaard (1986b) utilizou a distribuição estável positiva para a fragilidade. A partir das distribuições estáveis positivas, o autor derivou uma família de distribuições triparamétrica para a distribuição de fragilidade.

Segundo Silva (2001), na abordagem bayesiana dos modelos multiplicativos com fragilidade, a distribuição mais utilizada para a fragilidade também é a distribuição gama. Entretanto, outras distribuições de fragilidade têm sido investigadas em modelos multiplicativos bayesianos.

Considerando problemas de identificabilidade nos modelos de fragilidade multiplicativos, Manton et al. (1986) restringiram a função risco de base a uma constante ou à função risco Weibull. Os autores compararam modelos de mortalidade com fragilidade, utilizando funções de risco de base Weibull e Gompertz, juntamente com distribuições da fragilidade Gama e Gaussiana inversa e concluíram que o melhor ajuste foi obtido com o modelo de fragilidade Gama com função risco de base Weibull.

Elbers & Ridder (1982) apresentam condições para as quais as distribuições de fragilidade podem ser identificadas nos modelos de riscos proporcionais. Por exemplo, quando existe o valor médio da fragilidade,  $E(Z)$ , e pelo menos uma covariável contínua, Elbers & Ridder (1982) apresentaram a seguinte suposição de identificabilidade: a fragilidade  $Z$  é uma variável aleatória não negativa com função de distribuição  $F(Z)$  e  $E(Z) = 1$ .

### 2.6.2 Métodos de estimação

Métodos de estimação para o modelo de fragilidade Gama são mais complexos que para os modelos de riscos proporcionais porque, além de  $\beta$ , o parâmetro  $\eta$  da distribuição de fragilidade deve ser estimado.

Segundo Klein, (1992) como o modelo é formado por uma parte observada e outra não, o algoritmo EM, inicialmente proposto por Dempster et al. (1977), tem sido muito utilizado para estimar os parâmetros do modelo de fragilidade. Este algoritmo é caracterizado por um processo computacional iterativo para a obtenção de estimativas de máxima verossimilhança, entretanto, a convergência do algoritmo EM pode ser lenta.

Uma forma alternativa de estimação, sugerida por Therneau & Grambsch (2000), considera uma verossimilhança parcial penalizada, assim denominada porque uma função de penalidade, relacionada com a densidade da fragilidade, é incluída na log-verossimilhança parcial. Uma vantagem da maximização por verossimilhança parcial penalizada é que ela tende a reduzir o número de iterações quando comparada com o algoritmo EM.

Segundo Colossimo & Giolo (2006), procedimentos bayesianos que fazem uso de métodos computacionais intensivos, como o *Monte Carlo Markov chain* (MCMC), têm sido também sugeridos para estimação dos parâmetros dos modelos de fragilidade gama.

No modelo de fragilidade compartilhada,  $\lambda_0(t)$  pode não ser especificado (modelo semiparamétrico), como também pode ser assumida uma forma paramétrica específica para o risco basal (modelo paramétrico), sendo o processo de estimação para o modelo semiparamétrico, proposto por Ye et al. (2007) bem mais elaborado, quando comparado com o modelo paramétrico, segundo os próprios autores.

## **2.7 Classe geral de modelos com fragilidade multiplicativa**

Existem, atualmente, vários modelos e métodos de análise de eventos recorrentes (por exemplo, Hougaard, 2000; Therneau & Grambsch, 2000). No entanto, Peña & Hollander (2004) sugerem que ainda há a necessidade de uma classe de modelos mais geral e flexível que incorpore simultaneamente os efeitos de covariáveis ou variáveis concomitantes, o impacto da acumulação de tempos de ocorrências de eventos sobre a unidade ou indivíduo, o efeito de variáveis latentes ou não observadas que, para cada unidade, denota a correlação entre os tempos intereventos, bem como o efeito das intervenções realizadas após cada ocorrência de evento.

A maior parte das extensões do modelo de Cox (1972) lida com a maioria desses efeitos, mas não simultaneamente. Então, é necessária a utilização de um modelo mais geral que permita incorporar diferentes efeitos que afetam as ocorrências de eventos. Peña & Hollander (2004) propuseram uma nova classe de modelos com maior generalização do que a maioria dos atuais modelos de confiabilidade e também baseados no modelo de Cox (1972). Essa classe leva em conta, simultaneamente, efeitos de intervenção, efeitos do acúmulo de eventos recorrentes e efeitos de variáveis concomitantes e efeito das variáveis latentes.

Segundo Wilson et al. (2005), estudos de confiabilidade geralmente tratam modelos que lidam com sistemas reparáveis. Esses modelos incorporam a ideia de idade efetiva, também chamada idade virtual, a qual descreve o efeito da reparação ou intervenção. Talvez esses modelos sejam mais facilmente introduzidos em confiabilidade do que em problemas biomédicos, devido à capacidade de controlar a idade efetiva. A principal razão é que, ao fracasso, pode-se decidir se o sistema será substituído por um novo idêntico (reparação perfeita) ou se será restaurado ao seu estado pouco antes da falha (reparação mínima).

No entanto, quando a pesquisa trata de pacientes, não se pode decidir o seu grau de “reparação”. O objetivo, nesses casos, seria o de sempre, ou seja, eliminar todas as manifestações da doença (reparação perfeita). Por outro lado, nem sempre é obtida a mesma resposta, podendo-se observar uma remissão completa da doença, remissões parciais ou nenhuma resposta. Embora pareça natural que se considere o efeito das intervenções realizadas, nenhum dos principais modelos utilizados em estudos biomédicos incorpora esse efeito (González, 2006).

Exemplificando o efeito das intervenções, considere pessoas com doença cardíaca coronariana, às quais se recomendam alterar seu estilo de vida, reduzindo o nível de estresse, acabando com o hábito de fumar ou fazendo atividade física regularmente. Esses conselhos tentam modificar a probabilidade de se apresentar um novo ataque cardíaco. Considerando estudos de câncer, em alguns tumores, os médicos podem desempenhar algumas intervenções profiláticas ou curativas, como quimioterapia, radioterapia ou transplante de medula óssea do paciente para melhorar o status da doença.

Outra justificativa para se utilizar modelos baseados em idade efetiva é que a maioria dos modelos de sobrevivência para eventos recorrentes estudados na área de medicina e saúde pública baseia-se em extensões do modelo de Cox (1972). Assume-se, então, que o efeito depois de cada intervenção é sempre o mesmo. Por exemplo, modelos que empregam a formulação tempo total assumem que todas as intervenções produzem uma mínima melhora no paciente, ou seja, a doença continua de uma forma estável. Modelos baseados na formulação *gap time* assumem que todas as intervenções conduzem a uma perfeita recuperação do paciente, ou seja, a doença desaparece ou uma remissão completa é alcançada.

Muitas vezes, não é verdade que o efeito da intervenção no tratamento será sempre o mesmo. Por exemplo, pacientes com recorrência de tumores,

geralmente, são tratados depois de cada recidiva e obtêm uma resposta diferente após cada tratamento. Alguns pacientes alcançam uma remissão completa, outros, uma remissão mínima e outros entre nenhuma e remissão completa (Peña et al., 2003).

González (2006) afirma que, em virtude do modo tempo sequencial como os eventos recorrentes ocorrem, existe um aspecto adicional para esses estudos, que ainda não haviam sido considerados nos modelos existentes. Esse aspecto é a capacidade de realizar intervenções sobre o indivíduo após a ocorrência de cada evento. Por exemplo, quando um indivíduo abusa do álcool, intervenções na forma de métodos psicológicos e fisiológicos (hospitalização, terapia, elevação do nível de açúcar no sangue, hemoterapia, etc.) são realizadas. Similarmente, a um paciente que teve ataque cardíaco, seria recomendado que ele mudasse seu estilo de vida, isto é, mudança de hábitos alimentares, redução no nível de estresse, etc.

O objetivo primário de tais intervenções é prolongar, se não eliminar, a recorrência do evento. Sendo assim, tais intervenções podem ser vistas como a melhora do indivíduo ou unidade. A melhora seria possível porque, geralmente, as informações obtidas em cada ocorrência do evento podem ser utilizadas, bem como o conhecimento que é adquirido entre as ocorrências dos eventos, o que pode ajudar na formulação de novas estratégias de intervenção.

Deve-se levar em conta, porém, especialmente em estudos relacionados à medicina e à saúde pública, que normalmente lidam com seres humanos, que, apesar de a intervenção poder trazer melhorias, outros fatores, como o efeito de enfraquecimento sobre o acúmulo de ocorrências de eventos no indivíduo e os efeitos adversos do envelhecimento e de outras covariáveis tempo-dependentes, podem superar o efeito da intervenção. Assim, quando todos estes efeitos são considerados, segundo Peña & Hollander (2004), o tempo para a próxima ocorrência de evento pode ser ainda menor, em um sentido estocástico, em

relação ao tempo anterior à interocorrência. É, portanto, de grande importância que qualquer tentativa de modelagem leve em consideração os efeitos das intervenções simultaneamente com os efeitos de acumulação de eventos recorrentes e variáveis concomitantes relevantes.

González et al. (2005) utilizaram o modelo geral para eventos recorrentes proposto por Peña & Hollander (2004), no qual o efeito das intervenções é representado pela idade efetiva do processo sobre a função risco de base para tratar repetidas recidivas de câncer. Para acomodar a situação de recorrência do câncer, os autores propuseram uma função de idade efetiva que codifica três possíveis respostas terapêuticas: remissão completa, remissão parcial, e resposta nula. O modelo proposto também incorpora o efeito de covariáveis, o impacto das recidivas anteriores e a heterogeneidade entre os indivíduos. Nesse trabalho foram analisados os tempos de recidiva de 63 pacientes com um determinado subtipo de linfoma e os resultados foram comparados com aqueles obtidos utilizando outros métodos existentes na literatura.

Peña (2006) apresentou uma revisão na qual, é possível ter uma visão geral dos trabalhos recentes que envolvem a modelagem e a análise de eventos recorrentes, tanto na área de engenharia quanto de saúde pública e biomedicina, entre outras. O autor enfatiza a importância de incorporar o impacto do acúmulo de ocorrências de evento no modelo a ser utilizado, bem como os efeitos das intervenções realizadas após cada ocorrência de evento, simultaneamente com os efeitos das covariáveis e uma potencial associação entre os tempos interevento dentro de uma unidade. O autor apresenta métodos de inferência estatística para esta classe de modelos, ilustrados por meio de um conjunto de dados reais.

González (2006) fornece os procedimentos para estimar os parâmetros da classe geral de modelos para eventos recorrentes (Peña & Hollander, 2004). Uma aproximação baseada em inferência semiparamétrica foi desenvolvida.

Quando o modelo sem fragilidade foi estimado, uma generalização da verossimilhança parcial para o modelo de Cox foi obtida. Por outro lado, quando as fragilidades foram incluídas no modelo, o algoritmo do EM foi desenvolvido. O autor apresenta, ainda, outra metodologia, baseada em uma aproximação da verossimilhança penalizada e comprovou a validade da metodologia aplicada por meio de um estudo de simulação.

Peña et al. (2007) descrevem procedimentos para estimar os parâmetros da classe geral de modelos semiparamétricos para eventos recorrentes. Os procedimentos de estimação são semiparamétricos, já que a função risco de base é especificada não parametricamente. A estimação dos parâmetros é descrita para os dois casos, com e sem o termo de fragilidade. As propriedades da distribuição amostral dos estimadores são examinadas por meio de um estudo de simulação e as consequências de erros de especificação do modelo são analisadas. Os autores aplicaram tais procedimentos a um conjunto de dados relativos ao tempo até a reincidência de câncer da bexiga e os resultados da análise foram comparados com aqueles obtidos quando são utilizados outros três modelos para análise de dados de eventos recorrentes: o modelo proposto por Andersen & Gill (1982), o proposto por Wei et al. (1989) e o proposto por Prentice et al. (1981), que são extensões do modelo de Cox (1972).

Stocker & Peña (2007) consideraram a classe geral de modelos para eventos recorrentes sob uma especificação totalmente paramétrica da função risco de base, avaliando dois casos, aqueles em que o modelo incorpora ou não os termos de fragilidade. Os autores apresentaram estimadores para os parâmetros do modelo e suas propriedades, finitas e assintóticas foram verificadas. Para se ajustar a classe geral de modelos com uma função risco de base paramétrica, foi utilizada a distribuição Weibull e os procedimentos foram aplicados a um conjunto de dados sobre máquinas de subsistemas hidráulicos de mineração.

### 2.7.1 Formulação matemática

Seja  $\mathbf{X} = (X_1, \dots, X_T)^\top$ , em que  $t$  denota vetor/matriz transposta, um vetor de covariáveis (idade, raça, sexo, por exemplo) para determinada unidade em observação, podendo essas covariáveis ser ou não tempo-dependentes. Sejam  $S_1, S_2, \dots$  os sucessivos tempos de interocorrência de um evento (*gap times*) e  $T_1, T_2, \dots$ , os sucessivos tempos de calendário para ocorrência desse evento.

Então,

$$T_0 \equiv 0, T_1 = S_1, T_2 = S_1 + S_2, T_3 = S_1 + S_2 + S_3, \dots$$

Seja  $\mathcal{F}_t = \{\mathcal{F}_t; t \geq 0\}$  uma coleção crescente e contínua à direita de  $\sigma$ -álgebra para esta unidade, isto é, uma filtragem (Fleming & Harrington, 1991; Andersen et al., 1993). Em particular,  $\mathcal{F}_t$  contém informações sobre o número de tempos de ocorrência de eventos no intervalo de tempo  $[0, t]$ , informação sobre covariáveis e informações sobre os tipos de intervenções efetuadas sobre as ocorrências de evento.

Segundo Peña & Hollander (2004), o modelo probabilístico para as sucessivas ocorrências de um evento recorrente é uma especificação de uma coleção de funções de distribuições conjuntas de  $\{T_1, T_2, T_3, \dots\}$ .

Por causa da natureza dinâmica ou característica tempo sequencial do ajuste e como consequência das intervenções que são realizadas, essas especificações são facilitadas para reafirmar o modelo em termos de funções risco ou intensidades de falhas.

Seja  $\{N(t); t \geq 0\}$  o processo que representa o número de ocorrências de um evento recorrente de interesse durante o período  $[0, t]$ , e  $\{Y(t); t \geq 0\}$ . O processo de risco será  $Y(t) = 1$ , se a unidade ainda está sob observação um tempo  $t$  e  $Y(t) = 0$  se a unidade não está sob observação no tempo  $t$ . Seja  $\{A(t); t \geq 0\}$  um processo previsível não decrescente, tal que o processo  $\{M(t); t \geq 0\}$ , em que  $M(t) = N(t) - A(t)$ ,  $t \geq 0$  é um *martingale* local.

Assume-se, como em Andersen et al. (1993), que  $A(t) = \int_0^t d\Lambda(\omega)$ , em que  $\{\Lambda(t): t \geq 0\}$  é um processo previsível não decrescente, satisfazendo  $d\Lambda(\omega) = Y(\omega)\lambda(\omega)d\omega$ , em que  $\lambda(\omega): \omega \geq 0$  é um processo previsível não negativo (veja Aalen, 1978).

Segundo Jacod (1975), ele terá uma interpretação intuitiva e prática que, para  $h > 0$ , é suficientemente pequeno; a quantidade  $Y(t)\lambda(t)$  representa a probabilidade condicional aproximada, dado  $\mathcal{F}_{t-}$  dos eventos recorrentes ocorridos no intervalo  $[t, t+h)$ . Como indicado pela sua mensurabilidade com respeito a  $\mathcal{F}_{t-}$ , a intensidade do processo  $\lambda(t)$  pode depender das covariáveis da unidade e do número de ocorrências de eventos durante o período  $[0, t]$ .

Assuma, agora, a existência de um espaço de probabilidade completo  $(\Omega, \mathcal{F}, \mathbf{P})$  associado com a filtragem:

$$\mathbf{F} = \{\mathcal{F}_t: t \in [0, \tau]\}$$

em que  $0 \leq \tau \leq \infty$  é o limite superior do período de estudo.

Toda a realidade aleatória é definida em  $(\Omega, \mathcal{F})$ . Em particular, os tempos de interocorrência  $S_1, S_2, \dots$ , os tempos de calendário  $T_1, T_2, \dots$ , de ocorrências de eventos e os processos observáveis  $\{N(t): t \in [0, \tau]\}$  e  $\{Y(t): t \in [0, \tau]\}$  são definidos em  $(\Omega, \mathcal{F})$ . Essa proposta de nova classe de modelos postula que a taxa de intensidade do processo  $\{\lambda(t|X): t \in [0, \tau]\}$ , para um indivíduo ou unidade com vetor de covariáveis  $\mathbf{X} = (X_1, \dots, X_q)^t$ , as quais podem ser tempo-dependentes, é da forma:

$$\lambda(t|X) = \lambda_0[\varepsilon(t)]\rho[N_{t-}] \psi(\beta^t X) \quad (2.27)$$

em que  $\lambda_0(\cdot)$  é uma função risco de base desconhecida;  $\rho(\cdot)$  é uma função não decrescente ou não crescente de  $N = \{1, 2, \dots\}$  em  $\mathfrak{R}_+ = [0, \infty]$ , que pode depender de parâmetros desconhecidos com a condição  $\rho_0 = 1$ ;  $\psi(\cdot)$  é uma função não negativa de  $\mathfrak{R} - (-\infty, \infty)$  em  $\mathfrak{R}_+$  que tem forma conhecida (usualmente uma função exponencial);  $\beta = (\beta_1, \beta_2, \dots, \beta_q)^t$  é um vetor de

coeficientes de regressão desconhecidos e  $\{\varepsilon(t); t \in [0, \tau]\}$  é um processo previsível observável satisfazendo as condições:

- (i)  $\varepsilon(t) = \varepsilon_0$  quase certamente, em que  $\varepsilon_0$  é um numero real não negativo;
- (ii)  $\varepsilon(t) \geq 0$  quase certamente;
- (iii) Para  $t \in [T_{k-1}, T_k)$ ,  $\varepsilon(t)$  é quase certamente monótona e diferenciável com derivada positiva  $\varepsilon'(t) \in (0,1]$ .

Essa proposta modela a intervenção, tal como uma mudança na idade efetiva da unidade, diferente dos outros modelos encontrados na literatura, pois a maioria desses modelos incorpora tais efeitos no componente de regressão. Nesse caso, segundo Stocker & Peña (2007), se  $\varepsilon_1(t) = t$ , a reparação causada pela intervenção e medida pela idade efetiva do processo é dita mínima e se  $\varepsilon_1(t) = 1$ , a reparação é dita perfeita.

No modelo (2.27), a função  $\rho(\cdot)$  representa o efeito de acumulação de ocorrência de eventos. Segundo Gonzalez (2006), em muitas situações biomédicas, ela será assumida como uma função não decrescente de  $N_t^-$ , uma vez que é natural assumir que as ocorrências de evento tenham o efeito de enfraquecer a unidade ou indivíduo.

Esta característica não crescente, segundo Wilson et al. (2005), também é prevalente em modelos de confiabilidade em que em cada ocorrência de evento, falhas ou defeitos no sistema ou componente são eliminados, o que leva a melhorias. Assim, geralmente, exige-se que essa função seja monótona, quer não-decrescente ou não-crescente, dependendo do contexto ou situação.

A função  $\psi(\cdot)$  nesse modelo claramente serve ao objetivo de conter os efeitos das variáveis concomitantes. No modelo (2.27), os efeitos das intervenções, os efeitos da acumulação de ocorrência de eventos e os efeitos das covariáveis são, então, levados em conta simultaneamente. Além disso, segundo González (2006), existe uma interação entre esses efeitos, no sentido de que,

imediatamente após a intervenção, em um sentido total, a unidade não precisa sempre estar melhor, se comparada ao seu estado imediatamente antes da ocorrência do evento porque efeitos de melhoria podem ser compensados pelos outros dois efeitos.

Suponha  $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)$ , um vetor de variáveis aleatórias positivas independentes e identicamente distribuídas de uma distribuição paramétrica conhecida e que essas variáveis aleatórias são não observáveis e afetam a ocorrência de evento nos indivíduos. Além disso, seja  $\mathbf{F} = (\mathcal{F}_t : 0 \leq t \leq t^*)$  a filtragem ou a história em algum espaço de probabilidade  $(\Omega, \mathcal{F}, \mathcal{P})$ , tal que  $X_i$  e  $\mathcal{F}_t$  são preditíveis e tal que  $X_i$  são processos de contagem com relação a  $\mathcal{F}$ . Finalmente, seja,  $(\mathcal{E}_i(t) : 0 \leq t \leq t^*)$ ,  $i = 1, 2, \dots, n$ , a idade efetiva do processo.

A classe geral de modelos proposta por Peña & Hollander (2004) é obtida como segue.

Condicionalmente em  $\mathbf{Z}$ , o compensador de  $N_i^\dagger$  é  $A_i^\dagger(t|\mathbf{Z}, X_i) : 0 \leq t \leq t^*$  com

$$A_i^\dagger(t|\mathbf{Z}, X_i) = \int_0^t Y_i^\dagger(v) \lambda_i(v|\mathbf{Z}_t, X_i) dv \quad (2.28)$$

em que

$$\lambda_i(t|\mathbf{Z}, X_i) = Z_i \lambda_0[\mathcal{E}_i(t)] \rho[N_i^\dagger(t-); \alpha] \psi(\beta' X_i(t)) \quad (2.29)$$

A equação (2.29) é a função intensidade relacionada à classe geral de modelos para eventos recorrentes com fragilidade multiplicativa. Nela,  $\lambda_0$  é a função risco de base desconhecida. Os efeitos de acumulação de ocorrências de evento são codificados em  $\rho(\cdot; \alpha) : \mathbb{Z}_+ \equiv \{0, 1, 2, \dots\} \rightarrow \mathbb{R}_+$  que tem forma conhecida  $\rho(0; \alpha) = 1$ , uma vez que  $\rho(k; \alpha) = \alpha^k$  e com  $\alpha \in \mathcal{A} \subseteq \mathbb{R}^p$ . Se  $\alpha > 1$ , há uma indicação de que cada ocorrência de evento aumenta o risco de

eventos futuros. Caso contrário, se  $\alpha < 1$ , cada ocorrência de evento diminui o risco de eventos futuros e se  $\alpha = 1$  não existe nenhum efeito relacionado ao impacto da acumulação de tempos de ocorrências de eventos sobre os indivíduos.

Os efeitos das covariáveis são considerados em  $\psi(\cdot)$ , que é uma função não negativa com forma conhecida e  $\beta \in \mathcal{B} \subseteq \mathcal{R}^k$ . A dependência entre os tempos de ocorrência de eventos é modelada por  $Z_i$  fragilidades não observadas. O modelo também incorpora o efeito das intervenções realizadas,  $\varepsilon_i(\cdot)$ , por meio da função risco de base.

A identificabilidade estatística dessa classe de modelos sem fragilidades foi estabelecida no teorema 1 de Peña e Hollander (2004).

## 2.8 Inferência bayesiana

Segundo Paulino et al. (2003), a análise bayesiana de dados é uma alternativa importante em relação aos procedimentos clássicos de estimação e testes de hipóteses que vêm recebendo espaço crescente em aplicações nas mais diversas áreas.

A informação que se tem sobre uma quantidade de interesse  $\theta$  é fundamental na Estatística. O verdadeiro valor de  $\theta$  é desconhecido e a ideia é tentar reduzir esse desconhecimento. Além disso, a intensidade da incerteza a respeito de  $\theta$  pode assumir diferentes graus. Do ponto de vista bayesiano, esses diferentes graus de incerteza são representados por meio de modelos probabilísticos para  $\theta$ . Nesse contexto, segundo Ehlers (2007), é natural que diferentes pesquisadores possam ter diferentes graus de incerteza sobre  $\theta$  (especificando modelos distintos). Sendo assim, não existe nenhuma distinção entre quantidades observáveis e os parâmetros de um modelo estatístico; todos são considerados quantidades aleatórias.

Broemeling (1989) e Gelman et al. (2000) relatam que a característica essencial da inferência bayesiana está na utilização explícita de probabilidades para quantificar as incertezas que se têm sobre o parâmetro de interesse. Portanto, a inferência bayesiana é baseada no conceito de probabilidade subjetiva, que mede o grau de confiança que alguém deposita no acontecimento de um determinado evento do espaço amostral.

A inferência bayesiana é o processo de encontrar um modelo de probabilidade para um conjunto de dados e resumir o resultado por uma distribuição de probabilidade sobre os parâmetros do modelo e sobre quantidades não observadas, tais como predição para novas observações (Gelman et al., 2000).

De acordo com Paulino et al. (2003), a distribuição *a priori* é o único fator que diferencia a estatística clássica da bayesiana. Portanto, é alvo de crítica para muitos dos estatísticos clássicos, que alegam que a situação *a priori* é um processo subjetivo, já que dois pesquisadores podem ter diferentes graus de incertezas sobre uma quantidade desconhecida.

Em linhas gerais, segundo Savian (2009), a metodologia bayesiana consiste de informações referentes aos dados amostrais (função de verossimilhança), do conhecimento prévio a respeito dos parâmetros (distribuição *a priori*) e da obtenção da densidade *a posteriori* dos parâmetros por meio do teorema de Bayes. A utilização de informação *a priori* requer a especificação de uma distribuição para o parâmetro, devendo esta representar probabilisticamente o conhecimento que se tem sobre ele, antes da realização do experimento.

### **2.8.1 Distribuições *a priori***

Na inferência bayesiana, à incerteza sobre os parâmetros desconhecidos associa-se uma distribuição de probabilidade.

A distribuição *a priori* tem importante papel na análise bayesiana, sendo utilizada para descrever uma informação sobre os parâmetros desconhecidos, antes que se possam avaliar os dados em questão (Box & Tiao, 1992).

Segundo Pascoa (2006), a distribuição *a priori* é uma crença subjetiva e expressa o conhecimento que o pesquisador tem a respeito do problema, sendo, muitas vezes, baseada simplesmente na expectativa subjetiva do investigador, causando, assim, uma controvérsia pertinente de pesquisador para pesquisador.

O conhecimento prévio que o pesquisador tem sobre o parâmetro  $\theta$  é incorporado à análise por meio da densidade *a priori*  $\pi(\theta)$ , a qual deve representar a distribuição de probabilidade desse parâmetro antes da observação dos dados (Box & Tiao, 1992). O parâmetro pode ser um escalar ou um vetor de parâmetros  $\theta = (\theta_1, \theta_2, \dots, \theta_p)$ .

Os parâmetros indexadores das distribuições *a priori* são chamados de hiperparâmetros, distinguindo-os, assim, dos parâmetros de interesse. Inicialmente, os hiperparâmetros são considerados conhecidos e traduzem a informação que se tem sobre o parâmetro, antes da realização da amostra (Silva, 2008).

Uma classe importante de densidades *a priori* é a das chamadas *prioris* conjugadas. A ideia, segundo Ehlers (2007), é que as distribuições *a priori* e a *posteriori* pertençam à mesma classe de distribuições e, assim, a atualização do conhecimento que se tem de  $\theta$  envolve apenas uma mudança nos hiperparâmetros. Neste caso, o aspecto sequencial do método bayesiano pode ser explorado definindo-se apenas a regra de atualização dos hiperparâmetros, já que as distribuições permanecem as mesmas.

No entanto, conforme salientam Gelman et al. (2000), é essencial ter em mente que a distribuição deve adequadamente representar a incerteza *a priori* sobre os parâmetros, o que nem sempre ocorre com o uso de *prioris* conjugadas.

As distribuições *a priori* podem ser informativas ou não informativas. A partir do conhecimento prévio sobre o parâmetro  $\theta$ , pode-se descrever, aproximadamente, a distribuição de probabilidade que melhor o representa. Uma forma de expressar esse conhecimento, segundo Reis (2009), é utilizando informações fornecidas por pesquisas realizadas sobre um determinado assunto. Nesse caso, será utilizada uma *priori* informativa, descrevendo uma densidade  $\pi(\theta)$  que represente esse conhecimento.

Quando se tem pouca ou nenhuma informação sobre o parâmetro, pode-se usar uma *priori* não informativa. A utilização de uma distribuição *a priori* não informativa implica que a informação contida nos dados é dominante, no sentido de que o conhecimento *a priori* seja vago (Gelman et al., 2000). A ideia é pensar em todos os valores como igualmente prováveis, ou seja, com uma distribuição *a priori* uniforme ( $\pi(\theta) \propto k$ ) em que  $\theta$  é limitado em um intervalo real.

Segundo Box & Tiao (1992), a *priori* não-informativa não representa, necessariamente, o desconhecimento do pesquisador sobre o experimento, sendo utilizada também de forma a viabilizar a inferência *a posteriori*.

O argumento primeiramente invocado para gerar distribuições não informativas foi o princípio da razão insuficiente de Bayes-Laplace (Paulino et al., 2003). De acordo com este princípio, na ausência de razão suficiente para privilegiar uma possibilidade em detrimento de outras, decorrente da escassez de informação *a priori*, deve-se adotar a equiprobabilidade.

Quando o valor de um parâmetro é desconhecido, Jeffreys (1961) propõe uma regra para a escolha da distribuição *a priori*, a qual, segundo ele, atinge os casos mais comuns.

Seja uma observação  $X$  com função densidade de probabilidade  $\pi(\theta|X)$ . A *priori* não informativa de Jeffreys tem função de densidade dada por  $\pi(\theta) \propto [I(\theta)]^{1/2}$ , em que  $I(\theta)$  é a matriz de informação de Fisher. Se  $\theta$  for um vetor, então,  $\pi(\theta) \propto [\det I(\theta)]^{1/2}$ .

Algumas das maneiras de caracterizar a informação a *priori* são detalhadas por Berger (1985), Box & Tiao (1992) e Bernardo & Smith (1994).

A distribuição a *priori* do parâmetro  $\theta$ , que depende dos hiperparâmetros  $\phi$ , também pode ser especificada em estágios, sendo denominada, nesses casos, de *priori* hierárquica, denotada por  $\pi(\theta|\phi)$  em vez de  $\pi(\theta)$ . Neste caso, os hiperparâmetros não são considerados valores fixos, atribuindo-se também a eles uma distribuição a *priori*  $\pi(\phi)$ , especificando-se, assim, o segundo estágio na hierarquia.

Dessa forma, segundo Bernardo & Smith (1994), a distribuição a *priori* conjunta é simplesmente  $\pi(\theta, \phi) = \pi(\theta|\phi)\pi(\phi)$  e a distribuição marginal a *priori* de  $\theta$  pode, então, ser obtida por integração:

$$\pi(\theta) = \int \pi(\theta, \phi) d\phi = \int \pi(\theta|\phi)\pi(\phi) d\phi$$

Segundo Ehlers (2007), teoricamente, não há limitação quanto ao número de estágios, mas, devido às complexidades resultantes, as *prioris* hierárquicas são especificadas, em geral, em dois ou três estágios. Além disso, devido à dificuldade de interpretação dos hiperparâmetros em estágios mais altos, é prática comum especificar *prioris* não informativas para esses níveis.

### 2.8.2 Distribuição a *posteriori*

A análise bayesiana é conduzida combinando-se uma informação a *priori*  $\pi(\theta)$  com uma informação amostral  $\pi(x|\theta)$ . Segundo Paulino et al. (2003), a distribuição resultante dessa combinação  $\pi(\theta|x)$  é denominada distribuição a *posteriori* de  $\theta$  dado  $x$  e é dada por:

$$\pi(\theta|x) \propto \pi(\theta)\pi(x|\theta) \quad (2.30)$$

Segundo Box & Tiao (1992), a distribuição a *posteriori* pode ser vista como uma regra de atualização de  $\pi(\theta|x)$  pelo fato de que a distribuição a *priori*  $\pi(\theta)$  contém pouca informação sobre o parâmetro de interesse. Paulino et al.

(2003) salientam que, nesta regra de atualização, está implícito o princípio da verossimilhança, pois toda informação sobre o parâmetro está contida na distribuição amostral  $\pi(x|\theta)$ .

Uma vez que a distribuição a *posteriori* para os parâmetros está disponível, pode-se pensar em inferência. Em princípio, esta distribuição é a descrição mais completa possível de toda a informação de que se dispõe. Porém, para comparar e discutir resultados, bem como para resumir conclusões, é mais prático apresentar algumas quantidades que sintetizam esta distribuição (Gelman et al., 2000).

Segundo Savian (2009), um caso simples é a estimação pontual de  $\theta$ , em que se resume à distribuição marginal a *posteriori* por meio de um único valor,  $\hat{\theta}$ . É importante também associar alguma informação sobre o quão precisa é a especificação deste valor. As medidas de incerteza mais usuais, segundo Ehlers (2007), são: a variância e o coeficiente de variação para a média a *posteriori*, a medida de informação observada de Fisher para a moda a *posteriori* e a distância entre quartis para a mediana a *posteriori*.

### 2.8.3 Intervalo de credibilidade

Um resumo da distribuição a *posteriori* mais informativo do que qualquer estimativa pontual é obtido de uma região do espaço paramétrico  $\theta$ , que contenha uma parte substancial dessa distribuição (Paulino et al., 2003).

Assim, um intervalo  $(a, b)$  é chamado de intervalo de credibilidade  $100(1 - \alpha)\%$  para  $\theta$ , se

$$\int_a^b \pi(\theta|x) d\theta = 1 - \alpha$$

com  $0 \leq \alpha \leq 1$ .

Dada a infinidade de intervalos com o mesmo nível de credibilidade  $100(1 - \alpha)\%$ , interessa selecionar aquele com o menor comprimento possível

(Bolfarine & Sandoval, 2001). Os intervalos de comprimento mínimo são obtidos tomando-se os valores de  $\theta$  com maior densidade a *posteriori*. Essa região é denominada de HPD ou intervalo de máxima densidade a *posteriori*. Assim, quanto menor for o tamanho do intervalo, mais concentrada é a distribuição do parâmetro e a amplitude do intervalo informa sobre a dispersão de  $\theta$  (Reis, 2009).

Segundo Andrade (2008), para distribuições unimodais e simétricas, o intervalo de credibilidade e o HPD são idênticos. Havendo assimetria, os dois critérios produzem intervalos diferentes, sendo essa diferença dependente do grau de assimetria.

Os intervalos de credibilidade são úteis na análise de um teste de hipótese bayesiano. Suponha que o interesse seja testar  $H_0: \theta = \theta_0$  versus  $H_1: \theta \neq \theta_0$ . Para isso, primeiramente, constrói-se o intervalo de credibilidade para  $\theta$  e, se  $\theta_0$  está contido no intervalo,  $H_0$  não é rejeitada. Caso contrário, rejeita-se  $H_0$  (Bolfarine & Sandoval, 2001).

#### 2.8.4 Métodos computacionais

Segundo Paulino et al. (2003), para se inferir em relação a qualquer elemento de  $\theta$ , a distribuição conjunta a *posteriori* dos parâmetros deve ser integrada em relação a todos os outros parâmetros que a constituem, ou seja, procura-se obter a distribuição marginal de cada um dos parâmetros. A integração dessa distribuição, geralmente, não é analítica, sendo necessários algoritmos iterativos especializados, dentre os quais, os mais utilizados são os métodos Markov Chain Monte Carlo, ou MCMC.

Segundo Savian (2009), os métodos aproximados, baseados em simulação estocástica, consistem no tratamento de problemas reais a partir de reproduções em ambientes controlados pelo pesquisador. Alguns problemas apresentam componentes aleatórios, os quais não podem ser descritos de forma

exata e, sim, baseados em informações probabilísticas. Para esses problemas, o processo de simulação é estocástico, ou seja, baseado em distribuições de probabilidades.

Segundo Ehlers (2007), existem métodos de simulação não iterativos em que as amostras são obtidas da distribuição a *posteriori* em um único passo. São eles os métodos de Monte Carlo (ordinário e via função de importância) e os métodos de reamostragem (método da rejeição e método da reamostragem ponderada). Os métodos que fazem uso de técnicas de simulação iterativa são os métodos de Monte Carlo via cadeias de Markov (amostrador de Gibbs e o algoritmo de Metropolis-Hastings, por exemplo).

Em alguns casos, a expressão da condicional completa tem a forma de uma densidade conhecida e fácil de ser amostrada diretamente. Um método de simulação indicado nesses casos é o amostrador de Gibbs, um processo iterativo que gera valores que convergem para a densidade marginal, sem que se conheça a sua expressão (Gelfand, 2000).

Embora a distribuição condicional completa a *posteriori* possa ser obtida analiticamente, em determinadas situações, sua expressão não tem a forma de uma densidade conhecida ou fácil de ser amostrada. Em tais situações, outros métodos de simulação são indicados; entre eles, o algoritmo de Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970; Chib & Greenberg, 1995).

### 2.8.5 Métodos de Monte Carlo via cadeias de Markov

Uma cadeia de Markov é um processo estocástico no qual o próximo estado da cadeia,  $\phi_{t+1}$ , depende somente do estado atual,  $\phi_t$ , e dos dados, não da história passada da cadeia (Gamerman, 1997). Segundo este mesmo autor, as primeiras iterações são influenciadas pelo estado inicial,  $\phi_1$ , e são descartadas. Este período é conhecido como aquecimento da cadeia ou *burn-in*. Também se considera uma dependência entre as observações subsequentes da cadeia e, para

se obter uma amostra independente, as observações finais devem ser obtidas a cada  $k$  iterações, sendo este valor conhecido como salto, *thin* ou intervalo de amostragem.

A ideia básica desses métodos é a de transformar o problema estático em consideração, num problema de natureza dinâmica, construindo, para tal efeito, um processo estocástico temporal, artificial, que seja fácil de simular e que convirja para a distribuição marginal (Paulino et al., 2003).

Sendo assim, o objetivo, nesses casos, é obter uma amostra das distribuições marginais *a posteriori* dos parâmetros de interesse, por meio de um processo iterativo, utilizando as distribuições condicionais completas de cada parâmetro.

Por sua vez, esses valores gerados, segundo Reis (2009), são considerados amostras aleatórias de uma determinada distribuição de probabilidade, caracterizando, assim, o método de simulação Monte Carlo. Dessa forma, tem-se uma ação conjunta dos métodos, que resulta no processo MCMC, cujos principais algoritmos são o Metropolis- Hastings e o amostrador de Gibbs.

### **2.8.5.1 Metropolis-Hastings**

O algoritmo de Metropolis-Hastings é um método de Monte Carlo via cadeias de Markov (MCMC), definido a partir dos trabalhos de Metropolis et al. (1953) e Hastings (1970).

Quando as distribuições condicionais *a posteriori* não são facilmente identificadas como possuidoras de uma forma conhecida, não sendo possível gerar valores diretamente dessas distribuições, utiliza-se o algoritmo Metropolis-Hastings, o qual é estruturado a seguir:

Passo I – Inicializar o contador de iterações  $t = 0$  e especificar os valores iniciais  $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)})$ .

Passo II – Gerar um novo valor  $\theta^c$  da distribuição proposta  $q(\cdot | \theta_1)$ .

Passo III – Calcular a probabilidade de aceitação  $\alpha(\theta_1, \theta^c)$ .

$$\alpha(\theta_1, \theta^c) = \min\left(1, \frac{\pi(\theta^c | \theta_2, \dots, \theta_d) q(\theta_1 | \theta^c)}{\pi(\theta_1 | \theta_2, \dots, \theta_d) q(\theta^c | \theta_1)}\right)$$

Passo IV – Gerar  $u \sim U(0,1)$ .

Se  $u \leq \alpha$ , então, aceitar o novo valor e fazer  $\theta_1^{(t+1)} = \theta^c$ . Caso contrário, rejeitar o novo valor e fazer  $\theta_1^{(t+1)} = \theta_1$ .

Passo V – incrementar o contador de  $t$  para  $t + 1$  e voltar ao passo II, até atingir a convergência.

O algoritmo de Metropolis-Hastings permite gerar uma amostra da distribuição conjunta *a posteriori*  $\pi(\theta_1, \theta_2, \dots, \theta_d | x)$ , sendo  $d$  parâmetros, a partir das distribuições condicionais completas com formas desconhecidas. Tal algoritmo usa a ideia de que um valor é gerado de uma distribuição auxiliar ou candidata e este é aceito com uma dada probabilidade (Metropolis et al., 1953; Hastings, 1970).

Se essas distribuições possuem formas conhecidas, utiliza-se um caso especial do Metropolis-Hastings, o amostrador de Gibbs.

### 2.8.5.2 Amostrador de gibbs (*Gibbs Sampling*)

O amostrador de Gibbs é um método MCMC, utilizado para gerar variáveis aleatórias de uma distribuição sem utilizar diretamente sua densidade de probabilidade (Tomazella, 2003).

Esse método é um esquema iterativo de amostragem de uma cadeia de Markov, cujo núcleo de transição é formado pelas distribuições condicionais completas.

O amostrador de Gibbs foi inicialmente utilizado por Geman & Geman (1984) no contexto de restauração de imagens e, desde então, tem sido muito discutido e utilizado em diversas áreas de pesquisa (Gelfand & Smith, 1990; Gelfand, 2000).

Para descrever tal algoritmo de forma geral, considere o vetor aleatório  $\theta = (\theta_1, \theta_2, \dots, \theta_d)$  com distribuições condicionais completas para que os  $d$  parâmetros estejam disponíveis. Essas distribuições são denominadas por  $\pi(\theta_1 | \theta_2, \theta_3, \dots, \theta_d)$ ,  $\pi(\theta_2 | \theta_1, \theta_3, \dots, \theta_d)$ ,  $\dots$ ,  $\pi(\theta_d | \theta_1, \theta_2, \theta_3, \dots, \theta_{d-1})$ .

O algoritmo segue os seguintes passos:

Passo 0 – Especificar os valores iniciais  $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)})$  e inicializar o contador de iteração  $t = 1$ .

Passo I – Obter um novo valor  $\theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_d^{(t)})$  a partir de  $\theta^{(t-1)}$ , por meio de sucessivas gerações de valores:

$$\text{Gerar } \theta_1^{(t)} \sim \pi(\theta_1 | \theta_2^{(t-1)}, \theta_3^{(t-1)}, \dots, \theta_d^{(t-1)})$$

$$\text{Gerar } \theta_2^{(t)} \sim \pi(\theta_2 | \theta_1^{(t-1)}, \theta_3^{(t-1)}, \dots, \theta_d^{(t-1)})$$

⋮

$$\text{Gerar } \theta_d^{(t)} \sim \pi(\theta_d | \theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_{d-1}^{(t)})$$

Passo II – Incrementar  $t$  e voltar ao passo I, até atingir a convergência.

O número de iterações deve ser suficientemente grande para que a sequência de valores gerados se aproxime da distribuição de equilíbrio, ou seja, da densidade marginal desejada para cada parâmetro, quando se assume que a convergência foi atingida de acordo com Gelfand (2000).

### 2.8.6 Testes para avaliação de convergência

As técnicas MCMC constituem uma ótima ferramenta para a resolução de muitos problemas práticos na análise bayesiana, porém, apresentam algumas limitações. Entre elas, segundo Reis (2009), incluem-se as influências dos valores iniciais da cadeia, a correlação entre os dados gerados e a falta de convergência da cadeia.

Para a avaliação da convergência da cadeia de Markov para o estado de equilíbrio, vários testes foram propostos, dentre eles, os mais relevantes são as de Geweke (1992), Gelman & Rubin (1992), Heidelberg & Welch (1993) e Raftery-Lewis (1992). Esses testes estão implementados no pacote BOA (Bayesian Output Analysis) do *software* livre R (R Development Core Team, 2009).

O método de Geweke (1992) propõe o diagnóstico de convergência para cadeias de Markov, baseado no teste de igualdade de médias da primeira e da última parte da cadeia de Markov, geralmente dos primeiros 10% e dos últimos 50%. Ele se baseia em técnicas de análise espectral. Segundo Geweke (1992), a cadeia converge quando a maioria dos dados está entre os limites de uma distribuição normal padronizada.

O critério de Gelman & Rubin (1992) é um teste baseado na comparação entre duas ou mais cadeias paralelas. Ele analisa as variâncias dentro e entre as cadeias e utiliza essa informação para estimar o fator pelo qual o parâmetro escalar da distribuição marginal *a posteriori* deveria ser reduzido se a cadeia fosse repetida infinitas vezes. Este fator é expresso pelo valor  $\hat{R}$  (fator de redução de escala potencial ou fator de diagnóstico da convergência) e sugere que valores de  $\hat{R}$  próximos a 1 indicam que a convergência foi atingida para  $n$  iterações (Reis, 2009).

O método de Raftery & Lewis (1992) é baseado na acurácia de estimação do quantil. O método fornece as estimativas do *burn-in* (número de

iterações que devem ser descartadas), o número de iterações que devem ser computadas e a distância mínima ( $k$ ) de uma iteração à outra, *thin*, para se obter a subamostra aproximadamente independente. Outra saída importante, segundo Nogueira et al. (2004), é o fator de dependência, o qual é responsável pelo acréscimo multiplicativo ao número de iterações necessárias para se alcançar a convergência, sendo esta atingida quando esse valor é menor que 5.

O critério de Heidelberger & Welch (1993) utiliza testes estatísticos para avaliar a hipótese nula de estacionariedade da amostra gerada. Se a hipótese nula é rejeitada para um dado valor, o teste é repetido depois de descartadas os primeiros 10% das iterações. Se a hipótese é novamente rejeitada, outros 10% são descartados após o descarte dos 10% primeiros. Este processo é repetido até se ter uma proporção de 50% dos valores iniciais descartados. Se a hipótese for novamente rejeitada, ou o teste não conseguir ser realizado, isso indica falha da estacionariedade, implicando que é necessário um número maior de iterações. Se o teste for satisfatório, o número inicial de iterações descartadas é indicado como o tamanho do *burn-in* (Nogueira, 2004).

Segundo Reis (2009), outro critério utilizado para a verificação da convergência é a autocorrelação. De acordo com o autor, esse teste calcula a função de autocorrelação da cadeia com determinados valores de defasagens ou *lags*. Se a cadeia apresenta alta correlação, isto é um forte indicativo de falta da convergência.

A verificação informal da convergência utilizando técnicas gráficas é também um procedimento bastante útil. Segundo Gelman et al. (2000), os gráficos mais frequentes nesta análise são o de  $\theta$  ao longo das iterações e um gráfico da estimativa da distribuição a *posteriori* de  $\theta$ , por exemplo, um histograma. Problemas podem ser verificados com essas técnicas, pois a utilização de gráficos como única ferramenta pode não ser adequada. A escala dos gráficos pode fornecer uma falsa impressão de igualdade.

Nogueira et al. (2004), após realizarem intenso estudo de avaliação dos critérios de convergência, concluíram que, para uma avaliação precisa da convergência, deve-se seguir o seguinte procedimento:

1. aplicar Raftery & Lewis em uma amostra piloto e determinar o tamanho ideal da sequência;
2. determinar o tamanho do *burn-in* pelo critério de Heidelberger & Welch;
3. monitorar a convergência das sequências nas proximidades do tamanho ideal, indicado pelo critério de Raftery & Lewis, por meio dos critérios de Gelman & Rubin e Geweke.

### **2.8.7 Critérios para seleção de modelos**

Em aplicações práticas, em geral, existem incertezas com relação ao modelo que melhor se ajusta ao conjunto de dados, sendo necessário algum método que auxilie a tomada de decisão em favor do modelo mais plausível. Em inferência bayesiana, um dos critérios usados para esse fim é o Fator de Bayes (Jeffreys, 1961; Kass & Raftery, 1995).

Segundo Paulino et al. (2003), existem alguns critérios baseados em aproximações ao fator de Bayes entre dois modelos  $M_1$  e  $M_2$ , dentre os quais podem-se citar o critério de informação de Akaike (AIC) e o critério de informação bayesiano (BIC). Esses critérios levam em consideração a complexidade do modelo no critério de seleção. São critérios que, essencialmente, penalizam a verossimilhança, utilizando o número de variáveis do modelo e, eventualmente, o tamanho da amostra. Esta penalização é feita subtraindo-se do valor da verossimilhança uma determinada quantidade, que depende do quão complexo é o modelo (quanto mais parâmetros, mais complexo).

O critério de informação de Akaike (AIC), desenvolvido por Hirotugu Akaike sob o nome de “um critério de informação”, em 1971 e proposto em Akaike (1974), é uma medida relativa da qualidade de ajuste de um modelo estatístico estimado. Fundamenta-se no conceito de entropia, oferecendo uma medida relativa das informações perdidas, quando um determinado modelo é usado para descrever a realidade e é dado por:

$$AIC = -2 \lg L(\hat{\theta}) + 2(p) \quad (2.32)$$

em que  $p$  é o número de parâmetros.

Tendo-se um conjunto de dados e vários modelos concorrentes, pode-se classificá-los de acordo com o seu AIC, com aqueles com os menores valores de AIC sendo os melhores (Burnham & Anderson, 2004).

O critério de informação bayesiano, também chamado de critério de Schwarz, foi proposto por Schwarz (1978) e é um critério de avaliação de modelos definido em termos da probabilidade *a posteriori*, sendo assim chamado porque Schwarz o provou por meio de um argumento bayesiano.

O critério BIC é dado por:

$$BIC = -2 \log \left[ \frac{\sup_{M_1} \pi(x|\theta_1, M_1)}{\sup_{M_2} \pi(x|\theta_2, M_2)} \right] - 2(p_2 - p_1) \log n \quad (2.33)$$

A regra de decisão baseada no BIC consiste na escolha do modelo que apresentar menor valor de BIC.

### 3 MATERIAL E MÉTODOS

Suponha que  $n$  indivíduos tenham seus tempos de ocorrência de eventos observados por um período de tempo  $[0, T_i]$ . O tempo “zero” representa a entrada do indivíduo no estudo. O tempo  $T_i$  representa o tempo final de observação do  $i$ -ésimo indivíduo, sendo  $T_i$ , pré-fixado pelo pesquisador.

Para cada indivíduo, são registradas  $j = 1, 2, \dots, m_i$  observações, de modo que a última observação poderá ser censurada, caso não haja ocorrência de evento no exato tempo  $T_i$ . A ocorrência ou não de evento nos respectivos tempos de observações é dada por  $y_{ij}$ , de acordo com a TABELA 1. Caso haja a ocorrência de evento, então,  $y_{ij} = 1$ . Caso não haja ocorrência de evento, então,  $y_{ij} = 0$  e, nesse caso, a observação é dita censurada.

Neste trabalho, consideram-se três modelos para análise dos dados: a classe geral de modelos de fragilidade propostos por Peña & Hollander (2004), o modelo de fragilidade multiplicativo usual (Clayton, 1978), o qual é um caso particular da classe geral de modelos e o modelo de regressão de Cox (1972), o qual assume que os eventos ocorrem de modo independente, ou seja, que o tempo e a ocorrência de eventos não estão relacionados com a formação inicial e subsequentes ocorrências. As técnicas bayesianas serão utilizadas para comparar tais modelos. Para esta análise serão considerados os métodos de seleção de modelos critério de informação de Akaike (AIC) e critério de informação bayesiano (BIC).

Por simplicidade são apresentados primeiramente todos os procedimentos relacionados ao modelo de fragilidade multiplicativo e, em seguida, para o modelo geral para eventos recorrentes.

Para o modelo de Cox, considerado mais simples que os demais, pois não inclui a variável de fragilidade e assume que os acontecimentos ocorrem de

modo independente, a função de verossimilhança pode ser obtida substituindo-se a equação (2.21) em (2.6).

### 3.1 Modelo de fragilidade multiplicativo

Caso particular do modelo (2.29), considerando  $\rho[N_i^\dagger(t-); \alpha] = 1$  e  $\psi(\beta' X_i) = \exp(\beta' X_i)$ ;

$$\lambda_{ij}(t|Z, X_i) = Z_i \lambda_0[\mathcal{E}_i(t_{ij})] \exp(\beta' X_i). \quad (3.1)$$

Para indivíduos  $i = 1, 2, \dots, n$ , com vetor de covariáveis  $X_i$  e com  $\lambda_0(t) = \lambda_0(t; \theta)$ , função de verossimilhança para  $\beta, \theta$  é dada por:

$$L(\beta, \theta) = \prod_{i=1}^n \left[ \prod_{j=1}^{m_i} \lambda_i(t_{ij}, \theta)^{y_{ij}} \right] \exp\{-\Lambda_i(T_i, \theta)\} \quad (3.2)$$

em que  $y_{ij} = 1$ , em caso de falha e  $y_{ij} = 0$ , em caso de censura.  $\Lambda_i(T_i, \theta)$  é a função risco acumulada.

Sendo assim, para o indivíduo com fragilidade  $Z_i$ , tem-se que a função de verossimilhança condicional é como segue:

$$\begin{aligned} L(\beta, \theta | Z_i) &= \\ &= \prod_{i=1}^n \exp\{-Z_i \Lambda_0[\mathcal{E}_i(T_i), \theta] \exp(\beta' X_i)\} \left[ \prod_{j=1}^{m_i} [Z_i \lambda_0[\mathcal{E}_i(t_{ij}), \theta] \exp(\beta' X_i)]^{y_{ij}} \right] \\ &= \prod_{i=1}^n \left[ \exp\{-Z_i \Lambda_0[\mathcal{E}_i(T_i), \theta] \exp(\beta' X_i)\} (Z_i \exp(\beta' X_i))^{\sum_{j=1}^{m_i} y_{ij}} \times \right. \\ &\quad \left. \times \left[ \prod_{j=1}^{m_i} [\lambda_0[\mathcal{E}_i(t_{ij}), \theta]]^{y_{ij}} \right] \right] \end{aligned}$$

$$\begin{aligned}
&= \prod_{i=1}^n \left[ \prod_{j=1}^{m_i} \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{y_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \prod_{i=1}^n \exp\{-\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i)\} \\
&\times (\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i))^{\sum_{j=1}^{m_i} y_{ij}}.
\end{aligned}$$

Fazendo  $\sum_{j=1}^{m_i} y_{ij} = D_i$  ( $D_i$  conta o número de falhas do indivíduo  $i$ ),  
pode-se reescrever  $L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{Z}_i)$  como segue:

$$\begin{aligned}
&L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{Z}_i) \\
&= \prod_{i=1}^n \left[ \prod_{j=1}^{m_i} \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{y_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \prod_{i=1}^n \exp\{-\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i)\} \times \\
&\times (\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i))^{D_i}. \tag{3.3}
\end{aligned}$$

Assumindo  $Z_i \sim \text{Gama}\left(\frac{1}{\eta}, \frac{1}{\eta}\right)$ , em que  $\text{Var}(Z_i) = \eta$  e  $E(Z_i) = 1$ , a  
função de verossimilhança completa é dada por:

$$\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}_i) &= L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{Z}_i) \times \prod_{i=1}^n f(\mathbf{Z}_i, \boldsymbol{\theta}) \\
L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}_i) &= \\
&= \prod_{i=1}^n \left[ \prod_{j=1}^{m_i} \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{y_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \prod_{i=1}^n \exp\{-\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i)\} \\
&\times (\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i))^{D_i} \\
&\times \prod_{j=1}^{m_i} \left\{ \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} \exp\left\{-\mathbf{Z}_i \frac{1}{\eta}\right\} \mathbf{Z}_i^{\frac{1}{\eta}-1} \right\}. \tag{3.4}
\end{aligned}$$

Segundo Giolo (2003) e Matuda (2005), a estimação dos parâmetros do  
modelo de fragilidade, na versão paramétrica, é baseada na função de  
verossimilhança observada (ou marginal), isto é, na função de verossimilhança

não-condicionada, que se obtém ao integrar a função de verossimilhança completa, em relação a  $\mathbf{Z}$ .

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \int_0^{\infty} L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}_i) d\mathbf{Z}_i = \int_0^{\infty} L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{Z}_i) \prod_{i=1}^n f(\mathbf{Z}_i) d\mathbf{Z}_i$$

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \int_0^{\infty} \left\{ \prod_{i=1}^n \left[ \prod_{j=1}^{m_i} \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{\nu_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \prod_{i=1}^n \exp\{-\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i)\} \right. \\ \left. \times (\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i))^{D_i} \times \prod_{i=1}^n \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} \exp\left\{-\mathbf{Z}_i \frac{1}{\eta}\right\} \mathbf{Z}_i^{\frac{1}{\eta}-1} \right\} d\mathbf{Z}_i$$

Para o  $i$ -ésimo indivíduo, tem-se que:

$$L_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} (\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i))^{D_i} \prod_{j=1}^{m_i} \left[ \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{\nu_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \times \\ \times \int_0^{\infty} \{ \exp\{-\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i)\} (\mathbf{Z}_i)^{D_i} \\ \times \exp\left\{-\mathbf{Z}_i \frac{1}{\eta}\right\} \mathbf{Z}_i^{\frac{1}{\eta}-1} \} d\mathbf{Z} \\ = \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} (\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i))^{D_i} \prod_{j=1}^{m_i} \left[ \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{\nu_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \times \\ \times \int_0^{\infty} \left\{ \exp\left\{-\mathbf{Z}_i \left[\frac{1}{\eta} + \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' \mathbf{X}_i)\right]\right\} \mathbf{Z}_i^{\frac{1}{\eta} + D_i - 1} \right\} d\mathbf{Z}_i$$

Pode-se observar que esta integral em relação a  $\mathbf{Z}_i$  se refere a uma distribuição Gama  $\left(\frac{1}{\eta} + D_i, \frac{1}{\eta} + \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' X_i)\right)$ .

Então, a verossimilhança marginal para o  $i$ -ésimo indivíduo é dada por:

$$L_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}} \Gamma\left(\frac{1}{\eta} + D_i\right) (\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' X_i))^{D_i}}{\Gamma\left(\frac{1}{\eta}\right) \left[\frac{1}{\eta} + \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' X_i)\right]^{\frac{1}{\eta} + D_i}} \times \\ \times \prod_{j=1}^{m_i} \left[ \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{y_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right].$$

Logo, para  $n$  indivíduos,

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^n \left\{ \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}} \Gamma\left(\frac{1}{\eta} + \sum_{j=1}^{m_i} y_{ij}\right) (\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' X_i))^{D_i}}{\Gamma\left(\frac{1}{\eta}\right) \left[\frac{1}{\eta} + \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \exp(\boldsymbol{\beta}' X_i)\right]^{\frac{1}{\eta} + D_i}} \times \right. \\ \left. \times \prod_{j=1}^{m_i} \left[ \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{y_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \right\}. \quad (3.5)$$

### 3.2 Classe geral de modelos para eventos recorrentes com fragilidade multiplicativa

Considere-se o modelo (2.29), e o efeito da acumulação de eventos é dado por  $\alpha$ , em que  $\alpha$  conta o numero de falhas do indivíduo  $i$ .

A função de intensidade para o  $i$ -ésimo indivíduo é:

$$\lambda_{i|j}(t|\mathbf{Z}, X_i) = \mathbf{Z}_i \lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i). \quad (3.6)$$

Nesse caso, analogamente ao caso particular descrito na seção 3.1, a função de verossimilhança para  $\boldsymbol{\beta}, \boldsymbol{\theta}$  é como em (3.2). Sendo assim, para o

indivíduo com fragilidade  $\mathbf{Z}_i$ , tem-se que a função de verossimilhança condicional é como segue:

$$\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{Z}_i) &= \\
&= \prod_{i=1}^n \exp\{-\mathbf{Z}_i \Lambda_0[\boldsymbol{\varepsilon}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i)\} \left[ \prod_{j=1}^{m_i} [Z_i \lambda_0[\boldsymbol{\varepsilon}_i(t_{ij}), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i)]^{y_{ij}} \right] \\
&= \prod_{i=1}^n \exp\{-\mathbf{Z}_i \Lambda_0[\boldsymbol{\varepsilon}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i)\} (Z_i \exp(\boldsymbol{\beta}' X_i))^{D_i} \times \\
&\quad \times \left[ \prod_{j=1}^{m_i} [\lambda_0[\boldsymbol{\varepsilon}_i(t_{ij}), \boldsymbol{\theta}]]^{y_{ij}} \right] \\
&= \prod_{i=1}^n \prod_{j=1}^{m_i} \left[ \frac{(\lambda_0[\boldsymbol{\varepsilon}_i(t_{ij}), \boldsymbol{\theta}])^{y_{ij}}}{\Lambda_0[\boldsymbol{\varepsilon}_i(T_i), \boldsymbol{\theta}]} \right] \prod_{i=1}^n \exp\{-\mathbf{Z}_i \Lambda_0[\boldsymbol{\varepsilon}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i)\} \times \\
&\quad \times (Z_i \Lambda_0[\boldsymbol{\varepsilon}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i))^{D_i}.
\end{aligned}$$

A função de verossimilhança completa, assumindo  $\mathbf{Z}_i \sim \text{Gama}\left(\frac{1}{\eta}, \frac{1}{\eta}\right)$ , é dada por:

$$\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}_i) &= L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{Z}_i) \times \prod_{i=1}^n f(\mathbf{Z}_i) \\
L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}_i) &= \\
&= \prod_{i=1}^n \left[ \prod_{j=1}^{m_i} \frac{(\lambda_0[\boldsymbol{\varepsilon}_i(t_{ij}), \boldsymbol{\theta}])^{y_{ij}}}{\Lambda_0[\boldsymbol{\varepsilon}_i(T_i), \boldsymbol{\theta}]} \right] \prod_{i=1}^n \exp\{-\mathbf{Z}_i \Lambda_0[\boldsymbol{\varepsilon}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i)\} \\
&\quad \times (Z_i \Lambda_0[\boldsymbol{\varepsilon}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i))^{D_i}
\end{aligned}$$

$$\times \prod_{j=1}^m \left\{ \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} \exp\left\{-Z_i \frac{1}{\eta}\right\} Z_i^{\frac{1}{\eta}-1} \right\}. \quad (3.7)$$

A função de verossimilhança marginal é dada por:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \int_0^{\infty} L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{Z}_i) d\mathbf{Z}_i = \int_0^{\infty} L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{Z}_i) \prod_{i=1}^n f(\mathbf{Z}_i) d\mathbf{Z}_i \\ L(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \\ &= \int_0^{\infty} \left\{ \prod_{i=1}^n \left[ \prod_{j=1}^{m_i} \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{\gamma_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \prod_{i=1}^n \exp\{-\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' \mathbf{X}_i)\} \right. \\ &\quad \left. \times (\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' \mathbf{X}_i))^{D_i} \prod_{i=1}^n \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} \exp\left\{-Z_i \frac{1}{\eta}\right\} Z_i^{\frac{1}{\eta}-1} \right\} d\mathbf{Z}_i. \end{aligned}$$

Para o  $i$ -ésimo indivíduo, tem-se que:

$$\begin{aligned} L_i(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} (\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' \mathbf{X}_i))^{D_i} \prod_{j=1}^{m_i} \left[ \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{\gamma_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \times \\ &\quad \times \int_0^{\infty} \left\{ \exp\{-\mathbf{Z}_i \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' \mathbf{X}_i)\} (\mathbf{Z}_i)^{D_i} \exp\left\{-Z_i \frac{1}{\eta}\right\} Z_i^{\frac{1}{\eta}-1} \right\} d\mathbf{Z}_i \\ &= \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} (\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' \mathbf{X}_i))^{D_i} \prod_{j=1}^{m_i} \left[ \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{\gamma_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \times \\ &\quad \times \int_0^{\infty} \left\{ \exp\left\{-Z_i \left[\frac{1}{\eta} + \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' \mathbf{X}_i)\right]\right\} Z_i^{\frac{1}{\eta} + D_i - 1} \right\} d\mathbf{Z}_i. \end{aligned}$$

Pode-se observar que esta integral em relação a  $\mathbf{Z}_i$  se refere a uma distribuição Gama  $\left(\frac{1}{\eta} + D_i; \frac{1}{\eta} + \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i)\right)$ .

Então, a verossimilhança marginal para o indivíduo  $i$  é dada por:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} (\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i))^{D_i} \prod_{j=1}^{m_i} \left[ \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{\eta_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \times$$

$$\times \frac{\Gamma\left(\frac{1}{\eta} + D_i\right)}{\left[\frac{1}{\eta} + \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i)\right]^{\frac{1}{\eta} + D_i}}$$

Logo, para  $n$  indivíduos,

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^n \left\{ \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} (\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i))^{D_i} \prod_{j=1}^{m_i} \left[ \frac{(\lambda_0[\mathcal{E}_i(t_{ij}), \boldsymbol{\theta}])^{\eta_{ij}}}{\Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}]} \right] \right.$$

$$\left. \times \frac{\Gamma\left(\frac{1}{\eta} + D_i\right)}{\left[\frac{1}{\eta} + \Lambda_0[\mathcal{E}_i(T_i), \boldsymbol{\theta}] \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i)\right]^{\frac{1}{\eta} + D_i}} \right\}. \quad (3.8)$$

Considerando que a função risco de base tem uma distribuição Weibull dada por:

$$\lambda_0[\mathcal{E}(t)] = \delta(\mathcal{E}(t))^{\delta-1}, \quad (3.9)$$

tem-se uma abordagem totalmente paramétrica para o modelo Peña & Hollander (2004).

Levando em conta que a idade efetiva do processo utilizada é  $\mathcal{E}(t) = t$ , isto é, reparação mínima para a variável que mede o efeito das intervenções

sobre as ocorrências de evento, a função de verossimilhança para o modelo de Peña & Hollander (2004) é dada por:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^n \left\{ \prod_{j=1}^{m_i} \left[ \frac{[\delta(\mathcal{E}_i(t_{ij}))^{\delta-1}]^{\gamma_{ij}}}{T_i^\delta} \right] \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} (T_i^\delta \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i))^{D_i} \times \frac{\Gamma\left(\frac{1}{\eta} + D_i\right)}{\left[\frac{1}{\eta} + T_i^\delta \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i)\right]^{\frac{1}{\eta} + D_i}} \right\}. \quad (3.10)$$

### 3.3 Análise bayesiana para o modelo de fragilidade multiplicativo

Nesta seção é considerado o modelo de fragilidade multiplicativo, caso particular do modelo proposto por Peña & Hollander (2004), como descrito na seção 3.1. A análise bayesiana para o modelo de Cox, por simplicidade, não é descrita, sendo análoga à do modelo agora apresentado.

Sob a hipótese de independência, a densidade *a priori* conjunta para  $\boldsymbol{\beta}$ ,  $\delta$  e  $\eta$  é dada por:

$$\pi(\boldsymbol{\beta}, \delta, \eta) = \pi(\boldsymbol{\beta})\pi(\delta)\pi(\eta). \quad (3.11)$$

Combinando a função de verossimilhança (3.5) com a distribuição *a priori* (3.11), especificando a função risco de base como em (3.19) e a idade efetiva do processo como  $\mathcal{E}(t) = t$  a distribuição *a posteriori* conjunta para  $\boldsymbol{\beta}$ ,  $\delta$  e  $\eta$  é dada por:

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \delta, \eta | \mathbf{D}) &\propto \pi(\boldsymbol{\beta})\pi(\delta)\pi(\eta) \times \\
&\times \prod_{i=1}^n \left\{ \prod_{j=1}^{m_i} \left[ \frac{\left[ \delta (\mathcal{E}_i(\hat{t}_{ij}))^{\delta-1} \right]^{y_{ij}}}{T_i^\delta} \right] \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} (T_i^\delta \exp(\boldsymbol{\beta}' X_i))^{D_i} \times \right. \\
&\left. \times \frac{\Gamma\left(\frac{1}{\eta} + D_i\right)}{\left[\frac{1}{\eta} + T_i^\delta \exp(\boldsymbol{\beta}' X_i)\right]^{\frac{1}{\eta} + D_i}} \right\}, \tag{3.12}
\end{aligned}$$

As densidades a *posteriori* marginais dos parâmetros em (3.12) não são facilmente obtidas, pois a integração da densidade a *posteriori* conjunta não é trivial. Uma alternativa para a estimação dos parâmetros em situações como esta é a utilização dos métodos MCMC, descritos na seção 2.8.5.

A distribuição a *posteriori* condicional para  $\boldsymbol{\beta}$ , dado  $\delta$  e  $\eta$ , é dada por:

$$\pi(\boldsymbol{\beta} | \delta, \eta) \propto \prod_{i=1}^n \left\{ \frac{(\exp(\boldsymbol{\beta}' X_i))^{D_i}}{\left[\frac{1}{\eta} + T_i^\delta \exp(\boldsymbol{\beta}' X_i)\right]^{\frac{1}{\eta} + D_i}} \right\} \times \pi(\boldsymbol{\beta}). \tag{3.13}$$

A distribuição a *posteriori* condicional para  $\delta$ , dado  $\boldsymbol{\beta}$  e  $\eta$ , é dada por:

$$\begin{aligned}
\pi(\delta | \boldsymbol{\beta}, \eta) &\propto \pi(\delta) \times \\
&\times \prod_{i=1}^n \left\{ \prod_{j=1}^{m_i} \left[ \frac{\left[ \delta (\mathcal{E}_i(\hat{t}_{ij}))^{\delta-1} \right]^{y_{ij}}}{T_i^\delta} \right] \frac{(T_i^\delta)^{D_i}}{\left[\frac{1}{\eta} + T_i^\delta \exp(\boldsymbol{\beta}' X_i)\right]^{\frac{1}{\eta} + D_i}} \right\}. \tag{3.14}
\end{aligned}$$

A distribuição a *posteriori* condicional para  $\eta$ , dados  $\boldsymbol{\beta}$  e  $\delta$ , é dada por:

$$\pi(\eta | \boldsymbol{\beta}, \delta) \propto \prod_{i=1}^n \left\{ \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}} \Gamma\left(\frac{1}{\eta} + D_i\right)}{\Gamma\left(\frac{1}{\eta}\right) \left[\frac{1}{\eta} + T_i \delta \exp(\boldsymbol{\beta}' \mathbf{X}_i)\right]^{\frac{1}{\eta} + D_i}} \right\} \pi(\eta). \quad (3.15)$$

Como as distribuições a *posteriori* condicionais não têm forma conhecida, será utilizado o algoritmo Metropolis-Hastings, descrito na seção 2.8.5.1, para estimação dos parâmetros de interesse. Tal algoritmo permite simular observações de distribuições conjuntas a partir das densidades condicionais.

### 3.4 Análise bayesiana para a classe geral de modelos

Nesta seção será considerado o modelo proposto por Peña & Hollander (2004), como descrito na seção 3.2.

Sob a hipótese de independência, a densidade a *priori* conjunta para  $\alpha$ ,  $\boldsymbol{\beta}$ ,  $\delta$  e  $\eta$  é dada por:

$$\pi(\alpha, \boldsymbol{\beta}, \delta, \eta) = \pi(\alpha) \pi(\boldsymbol{\beta}) \pi(\delta) \pi(\eta). \quad (3.16)$$

Combinando a função de verossimilhança (3.10) com a distribuição a *priori* (3.16), a distribuição a *posteriori* conjunta para  $\boldsymbol{\beta}$ ,  $\delta$  e  $\eta$  é dada por:

$$\begin{aligned}
\pi(\alpha, \beta, \delta, \eta | \mathbf{D}) &\propto \pi(\alpha) \pi(\beta) \pi(\delta) \pi(\eta) \times \\
&\times \prod_{i=1}^n \left\{ \prod_{j=1}^{m_i} \left[ \frac{[\delta (\varepsilon_i(\hat{t}_{ij}))^{\delta-1}]^{\eta_{ij}}}{T_i^\delta} \right] \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}}}{\Gamma\left(\frac{1}{\eta}\right)} (T_i^\delta \alpha^{D_i} \exp(\beta' X_i))^{D_i} \times \right. \\
&\left. \times \frac{\Gamma\left(\frac{1}{\eta} + D_i\right)}{\left[\frac{1}{\eta} + T_i^\delta \alpha^{D_i} \exp(\beta' X_i)\right]^{\frac{1}{\eta} + D_i}} \right\}. \quad (3.17)
\end{aligned}$$

As densidades a *posteriori* marginais dos parâmetros em (3.17) também não são facilmente obtidas. Sendo assim, para a estimação dos parâmetros foi utilizado o método MCMC. Para a utilização do mesmo, como visto na seção anterior, consideram-se as distribuições a *posteriori* condicionais completas para os parâmetros de interesse.

A distribuição a *posteriori* condicional para  $\alpha$ , dado  $\beta$ ,  $\delta$  e  $\eta$ , é dada por:

$$\begin{aligned}
\pi(\alpha | \beta, \delta, \eta) &\propto \\
&\propto \prod_{i=1}^n \left\{ \frac{(\alpha^{D_i})^{D_i}}{\left[\frac{1}{\eta} + T_i^\delta \alpha^{D_i} \exp(\beta' X_i)\right]^{\frac{1}{\eta} + D_i}} \right\} \pi(\alpha). \quad (3.18)
\end{aligned}$$

A distribuição a *posteriori* condicional para  $\beta$ , dados  $\alpha$ ,  $\delta$  e  $\eta$ , é dada por:

$$\pi(\boldsymbol{\beta}|\alpha, \delta, \eta) \propto \prod_{i=1}^n \left\{ \prod_{j=1}^{m_i} \frac{(\exp(\boldsymbol{\beta}' X_{ij}))^{D_{ij}}}{\left[ \frac{1}{\eta} + T_i^\delta \alpha^{D_{ij}} \exp(\boldsymbol{\beta}' X_{ij}) \right]^{\frac{1}{\eta} + D_{ij}}} \right\} \times \pi(\boldsymbol{\beta}). \quad (3.19)$$

A distribuição a *posteriori* condicional para  $\delta$ , dados  $\alpha$ ,  $\boldsymbol{\beta}$  e  $\eta$ , é dada por:

$$\pi(\delta|\alpha, \boldsymbol{\beta}, \eta) \propto \pi(\delta) \times \prod_{i=1}^n \left\{ \prod_{j=1}^{m_i} \left[ \frac{[\delta (\varepsilon_i(\hat{t}_{ij}))^{\delta-1}]^{D_{ij}}}{T_i^\delta} \right] \frac{(T_i^\delta)^{D_{ij}}}{\left[ \frac{1}{\eta} + T_i^\delta \alpha^{D_{ij}} \exp(\boldsymbol{\beta}' X_{ij}) \right]^{\frac{1}{\eta} + D_{ij}}} \right\}. \quad (3.20)$$

A distribuição a *posteriori* condicional para  $\eta$ , dado  $\alpha$ ,  $\boldsymbol{\beta}$  e  $\delta$  é dada por:

$$\pi(\eta|\alpha, \boldsymbol{\beta}, \delta) \propto \prod_{i=1}^n \left\{ \frac{\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}} \Gamma\left(\frac{1}{\eta} + D_i\right)}{\Gamma\left(\frac{1}{\eta}\right) \left[ \frac{1}{\eta} + T_i^\delta \alpha^{D_i} \exp(\boldsymbol{\beta}' X_i) \right]^{\frac{1}{\eta} + D_i}} \right\} \pi(\eta). \quad (3.21)$$

Analogamente à seção 3.4.1, as distribuições a *posteriori* condicionais não têm forma conhecida. Assim, é utilizado o algoritmo Metropolis-Hastings, para a estimação dos parâmetros de interesse.

### 3.5 Inferência a *posteriori* para as fragilidades

Em estudos para analisar tempos de eventos recorrentes, é natural o interesse em se estimar as fragilidades de cada indivíduo.

Considerando que existem dificuldades para estimar as fragilidades individuais para os modelos de fragilidade multiplicativos e aditivos sob o ponto

de vista frequentista, contornam-se essas dificuldades utilizando-se as técnicas bayesianas para a análise dos parâmetros do modelo (Tomazella, 2003).

Assim, a distribuição a *posteriori* condicional para cada  $Z_i$ , dados  $(\beta, \theta)$ , pode ser obtida considerando a função de verossimilhança (3.4), que corresponde ao caso particular do modelo Peña & Hollander (modelo de fragilidade multiplicativo), em que a distribuição a *priori* para as variáveis de fragilidade  $Z_i$ 's é uma distribuição  $Gama\left(\frac{1}{\eta}, \frac{1}{\eta}\right)$ .

Logo,

$$\pi(Z_i | \beta, \theta) \propto Z_i^{\frac{1}{\eta} - D_i - 1} \exp\left\{-Z_i \left[\frac{1}{\eta} + \Lambda_0[\mathcal{E}_i(T_i, \theta)] \exp(\beta' X_i)\right]\right\}$$

então,

$$(Z_i | \beta, \theta) \sim Gama\left(D_i + \frac{1}{\eta}, \Lambda_0[\mathcal{E}_i(T_i, \theta)] \exp(\beta' X_i) + \frac{1}{\eta}\right) \quad i = 1, 2, \dots, n.$$

Neste caso,  $Z_i$  pode ser estimado via média a *posteriori* da distribuição de  $Z_i$ , dados  $(\beta, \theta)$ .

Logo,

$$\hat{Z}_i = \frac{D_i + \hat{\eta}^{-1}}{\hat{\eta}^{-1} + \Lambda_0[\mathcal{E}_i(T_i, \hat{\theta})] \exp(\hat{\beta}' X_i)} \quad (3.22)$$

em que as estimativas  $\hat{\eta}$ ,  $\hat{\theta}$ ,  $\hat{\beta}$  são obtidas por meio de amostras geradas via MCMC, como descrito na seção 2.8.5.

Considerando a função risco de base de uma distribuição Weibull, como em (3.9), as estimativas dos  $Z_i$ 's são dadas por

$$\hat{Z}_i = \frac{D_i + \hat{\eta}^{-1}}{\hat{\eta}^{-1} + T_i^\delta \exp(\hat{\beta}' X_i)}, \quad (3.23)$$

Para o caso geral do modelo Peña & Hollander (2004), a distribuição a *posteriori* condicional para cada  $Z_i$ , dados  $(\beta, \theta)$ , pode ser obtida considerando

a função de verossimilhança (3.7) em que a distribuição a *priori* para as variáveis de fragilidade  $\mathbf{z}_i$ 's é uma distribuição  $Gama\left(\frac{1}{\eta}, \frac{1}{\eta}\right)$ .

Assim, analogamente ao que foi feito para o caso particular do modelo descrito anteriormente, tem-se que:

$$\pi(\mathbf{Z}_i | \boldsymbol{\beta}, \boldsymbol{\theta}) \propto \mathbf{Z}_i^{\frac{1}{\eta} + D_i - 1} \exp\left\{-\mathbf{Z}_i \left[\frac{1}{\eta} + \Lambda_0[\mathcal{E}_i(T_i, \boldsymbol{\theta})] \alpha^{D_i} \exp(\boldsymbol{\beta}' \mathbf{X}_i)\right]\right\}$$

e

$$(\mathbf{Z}_i | \boldsymbol{\beta}, \boldsymbol{\theta}) \sim Gama\left(D_i + \frac{1}{\eta}, \Lambda_0[\mathcal{E}_i(T_i, \boldsymbol{\theta})] \alpha^{D_i} \exp(\boldsymbol{\beta}' \mathbf{X}_i) + \frac{1}{\eta}\right)$$

$$i = 1, 2, \dots, n.$$

Então,  $\mathbf{Z}_i$  pode ser estimado via média a *posteriori* da distribuição de  $\mathbf{Z}_i$  dados  $(\boldsymbol{\beta}, \boldsymbol{\theta})$ . Considerando a função risco de base de uma distribuição Weibull, como em (3.9), as estimativas dos  $\mathbf{z}_i$ 's são dadas por

$$\hat{\mathbf{Z}}_i = \frac{D_i + \hat{\eta}^{-1}}{\hat{\eta}^{-1} + T_i \hat{\delta} \alpha^{D_i} \exp(\hat{\boldsymbol{\beta}}' \mathbf{X}_i)} \quad (3.24)$$

e as estimativas  $\hat{\eta}$ ,  $\hat{\delta}$ ,  $\hat{\boldsymbol{\beta}}$ ,  $\hat{\sigma}$  são obtidas por meio de amostras geradas via MCMC.

### 3.6 Aplicação

Os dados em análise foram coletados exclusivamente para a aplicação neste trabalho e referem-se a 149 pacientes com insuficiência renal crônica (IRC), tratados no Centro de Terapia Renal Substitutiva da Santa Casa de Misericórdia de Lavras, Minas Gerais, no período de janeiro de 1994 a junho de 2008.

Neste estudo foi estipulado um tempo final de observação de 60 meses para cada indivíduo, desde o início do seu tratamento. O tratamento hemodialítico, segundo Gouvêa et al. (2009), é realizado, geralmente, três vezes

por semana, em sessões com duração média de três a quatro horas, com o auxílio de uma máquina, dentro de clínicas especializadas no tratamento de hemodiálise.

No período pré-estipulado, foram observadas as recorrências de eventos; após os 60 meses de acompanhamento dos pacientes, as observações passam a ser consideradas como censuradas.

Os eventos recorrentes de interesse neste estudo são as hospitalizações a que cada um desses pacientes foi submetido ao longo do período de observação. Durante esse período, são registrados os tempos em que ocorreram tais hospitalizações, em meses. Por exemplo, para o indivíduo 1, que pode sofrer até  $m_1$  hospitalizações, são registrados os tempos, em meses, da seguinte maneira:

$$t_{11} < t_{12} < t_{13} < \dots < t_{1m_1}$$

De cada paciente foram obtidas as seguintes informações (TABELA 8, no Anexo E) que constituem as covariáveis em estudo:

1. sexo: masculino ou feminino;
2. estado civil: casado, solteiro, viúvo, separado;
3. raça: dividida em dois grupos, branca ou parda e negra;
4. idade no início do tratamento (em anos);
5. procedência: dividida em dois grupos, microrregião de Lavras e outras regiões;
6. nefrite: (doença também chamada glomerulonefrite, é um termo utilizado para descrever enfermidades renais, nas quais a parte filtrante do rim, glomérulo, está inflamada) dividida em dois grupos: os que possuem e os que não possuem a doença;
7. diabetes *mellitus*: (moléstia caracterizada por distúrbios no metabolismo de açúcares, gorduras e proteínas que levam à falta de secreção da insulina, ao aumento da glicose no sangue e ao comprometimento de

- vários órgãos, destacando-se os rins) dividida em dois grupos, os que possuem e os que não possuem a doença;
8. hipertensão: (mais popularmente conhecida como "pressão alta", está relacionada com a força que o coração tem que fazer para impulsionar o sangue para todo o corpo) dividida em dois grupos, os que possuem e os que não possuem a doença;
  9. presença de rim policístico: (distúrbio hereditário no qual ocorre a formação de muitos cistos em ambos os rins. Os rins aumentam de volume, mas possuem menos tecido renal funcional) dividida em dois grupos, os que possuem e os que não possuem a doença;
  10. outras doenças (outras etiologias identificadas no início do tratamento, por meio de avaliação médica, considerada a doença básica que levou o paciente a um quadro de insuficiência renal) dividida em dois grupos, os que possuem e os que não possuem a doença.

Sendo assim, o vetor de covariáveis, seguindo numeração de 1 a 10 é

$$X' = (\text{sexo}, \text{estado civil}, \text{raça}, \text{idade}, \dots, \text{outras doenças})$$

Em que existem 11 parâmetros de regressão a serem estimados, associados as covariáveis em estudo. O vetor de parâmetros é  $\beta' = (\beta_0, \beta_1, \beta_2, \dots, \beta_{10})$ .

Considerando o modelo (2.29) e especificando a função risco de base como em (3.9), para o conjunto de dados apresentados, a função de intensidade é dada por:

$$\lambda_{i,j}(t|Z, X) = Z_i \delta t^{\delta-1} \alpha^{\beta_i} \exp(\beta_0 + \beta_1 \text{sexo} + \beta_2 \text{estado civil} + \dots + \beta_{10} \text{outras doenças}).$$

As características das distribuições a *posteriori* dos parâmetros dos modelos (3.1) e (3.6) com distribuição de fragilidade gama foram estimadas de amostras geradas pela técnica de Metropolis-Hastings, no *software* R (R Development Core Team, 2009). Foram geradas duas cadeias de 30.000

iterações para os parâmetros. As primeiras 5.000 iterações foram descartadas (*burn-in*) e, para assegurar a independência da amostra, considerou-se um espaçamento entre os pontos amostrados, de tamanho 10 (*thin*). Sendo assim, obteve-se uma amostra final de 2.500 observações para cada parâmetro em cada amostra.

Para a utilização do algoritmo Metropolis-Hastings, foram utilizadas, para gerar valores candidatos, as próprias densidades a *priori*. Segundo Tomazella (2003), esse procedimento é artificial, mas pode ser aplicado juntamente com o monitoramento da saída do algoritmo Metropolis-Hastings por meio da taxa de aceitação. Foi aplicado também o critério de convergência de Gelman e Rubin e o número de iterações foi considerado suficiente para a convergência (valores de  $\hat{R}$  próximos a 1 indicam que a convergência foi atingida para  $n$  iterações).

Existem muitas possibilidades para a escolha das distribuições a *priori*. Neste trabalho, a escolha da distribuição foi feita de acordo com o domínio no espaço paramétrico. Para o parâmetro  $\alpha$ , foi assumida a *priori*  $\alpha \sim \text{Gama}(10; 0,1)$ . Para o vetor de parâmetros  $\beta$ , foi assumida uma distribuição Normal multivariada com média 0 e variância 10,  $\pi(\beta) = N(\mu, \Sigma)$ , sendo  $\mu$  um vetor de médias de dimensão 1x11 e  $\Sigma$  uma matriz diagonal de dimensão 11x11. Para o parâmetro  $\delta$ , foi assumida  $\delta \sim \text{Gama}(10; 0,1)$  e para  $\eta$ , uma distribuição gama,  $\eta \sim \text{gama}(1; 1)$ .

Foi realizado um estudo de sensibilidade considerando outros valores para os hiperparâmetros das distribuições a *priori*, em que as diferentes escolhas levaram aos mesmos resultados.

Os programas desenvolvidos na implementação computacional dos exemplos de aplicação para os modelos propostos foram realizados por meio do *software* livre R (R Development Core Team, 2009) e podem ser obtidas diretamente com a autora desta tese.

## 4 RESULTADOS E DISCUSSÃO

Nesta seção são apresentados e discutidos os resultados relativos ao modelo geral (Peña & Hollander, 2004), ao modelo de fragilidade multiplicativo (Clayton, 1978) e ao modelo de regressão de Cox (Cox, 1972).

A suposição de riscos proporcionais para a utilização do modelo de regressão de Cox foi verificada por meio dos gráficos do logaritmo da função risco acumulado versus o tempo para todas as covariáveis. Os gráficos para algumas dessas covariáveis são apresentados no Anexo D.

### 4.1 Classe geral de modelos para eventos recorrentes com fragilidade multiplicativa

Na TABELA 2 são apresentadas as estimativas da média, desvio padrão, intervalo de credibilidade 95%, valores de  $\hat{R}$  para a verificação da convergência da cadeia por meio do critério de Gelman e Rubin e taxa de aceitação para os parâmetros do modelo.

Os gráficos do traço da cadeia e da densidade para os parâmetros  $\beta_1$  referente à covariável sexo e  $\beta_2$ , referente à covariável estado civil, são apresentados nas FIGURAS de 3 e 4. Os gráficos relacionados aos demais parâmetros do vetor  $\beta$  são apresentados no Anexo A.

TABELA 2 Resumo da distribuição *posteriori* do modelo geral para eventos recorrentes. Média, desvio padrão (DP), intervalo de credibilidade 95% (IC), teste de diagnóstico (R) e taxa de aceitação (TA) para os parâmetros do modelo.

Parâmetros	Média	DP	IC	R	TA
$\beta_0$	-2,76	0,22	[-3,08; -2,36]	1,01	73,98
$\beta_1$	-0,84	0,11	[-1,06; -0,68]	1,01	73,98

$\beta_2$	1,51	0,62	[0,71; 2,64]	1,01	73,98
$\beta_3$	0,63	0,23	[0,25; 1,00]	1,01	73,98
$\beta_4$	0,03	0,01	[0,02; 0,04]	1,01	73,98
$\beta_5$	0,01	0,01	[-0,01; 0,02]	1,01	73,98
$\beta_6$	0,16	0,13	[0,01; 0,41]	1,01	73,98
$\beta_7$	-2,20	0,3	[-2,76; -1,80]	1,01	73,98
$\beta_8$	-1,76	0,16	[-2,02; -1,36]	1,01	73,98
$\beta_9$	-2,88	0,20	[-3,16; -2,44]	1,01	73,98
$\beta_{10}$	-1,90	0,35	[-2,66; -1,45]	1,01	73,98
$\sigma$	0,37	0,02	[0,32; 0,40]	0,99	49,3
$\eta$	1,23	0,57	[0,50; 2,32]	1,01	76,44
$\alpha$	1,80	0,18	[1,49; 2,19]	0,99	54,26

Com base nos resultados observados na TABELA 2, pode-se concluir que as covariáveis que estão associadas com um elevado risco de re-internação são estado civil ( $\beta_2$ ), raça ( $\beta_3$ ), idade ( $\beta_4$ ) e doença nefrite ( $\beta_5$ ), pois apresentam valores positivos para a média. O risco de reinternações de pacientes do sexo masculino é menor do que o risco de pacientes femininos. Pacientes solteiros/viúvos/separados têm 4,52 vezes mais risco de reinternações do que pacientes casados. Pacientes negros têm risco de reinternação 1,89 vez maior que brancos/pardos. Pacientes com nefrite têm risco de reinternação 1,17 vez maior do que os pacientes que não possuem tal doença.

Como se pode observar, a única covariável que não é estatisticamente significativa é a procedência. A significância estatística, neste caso, pode ser verificada por meio dos intervalos de credibilidade para os coeficientes estimados das covariáveis. Se o intervalo contiver o valor “zero”, a covariável é considerada não-significativa.

O valor estimado para a variância da distribuição de fragilidade dado por  $\eta = 1,23$  e o intervalo de credibilidade  $[0,50;2,32]$  indicam que existe uma heterogeneidade não observada significativa entre os indivíduos.

Considerando que  $\alpha = 1,80$  é significativo no modelo e maior que 1, há uma indicação de que cada internação aumenta o risco de hospitalizações futuras.

A convergência das cadeias de todos os parâmetros do modelo foi verificada por meio do critério de Gelman & Rubin, em que todos os casos apresentaram valores de  $\hat{R}$  próximos a 1 e também por meio da visualização gráfica do traço apresentado nas FIGURAS 3- 7.

Os valores das taxas de aceitação apresentaram resultados satisfatórios, como se pode observar na TABELA 2.

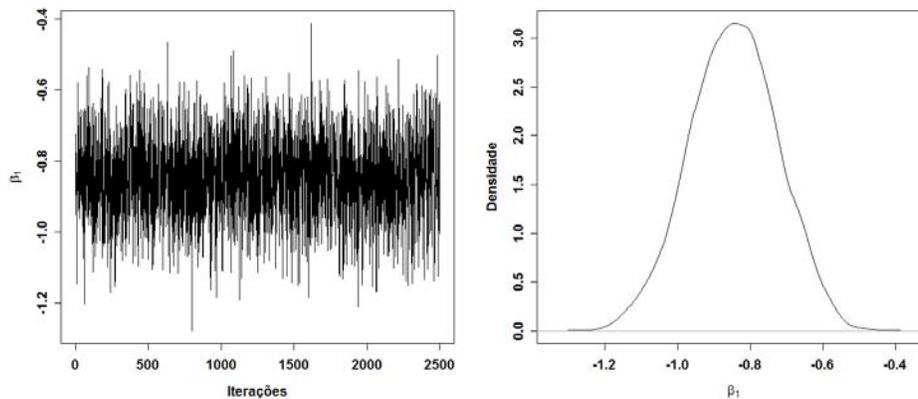


FIGURA 3 Traço da cadeia gerada pelo método MCMC e densidade *a posteriori* para o parâmetro  $\beta_1$ , referente à covariável sexo.

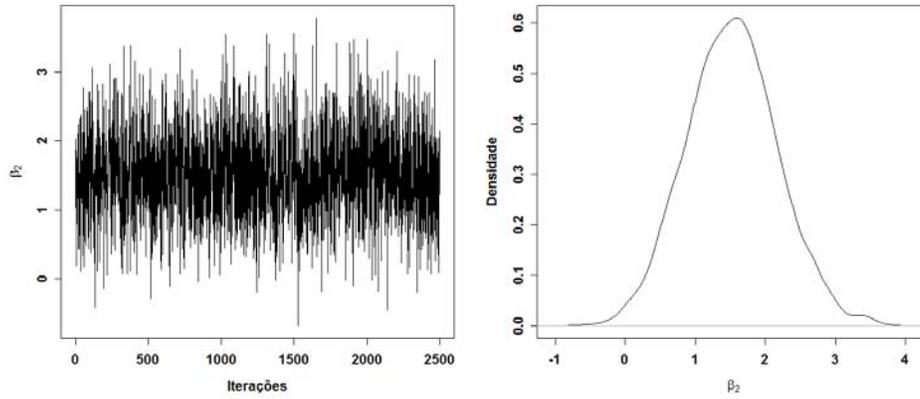


FIGURA 4 Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_2$ , referente à covariável estado civil.

Os gráficos do traço da cadeia e da densidade para o parâmetro  $\alpha$  são apresentados na FIGURA 5.

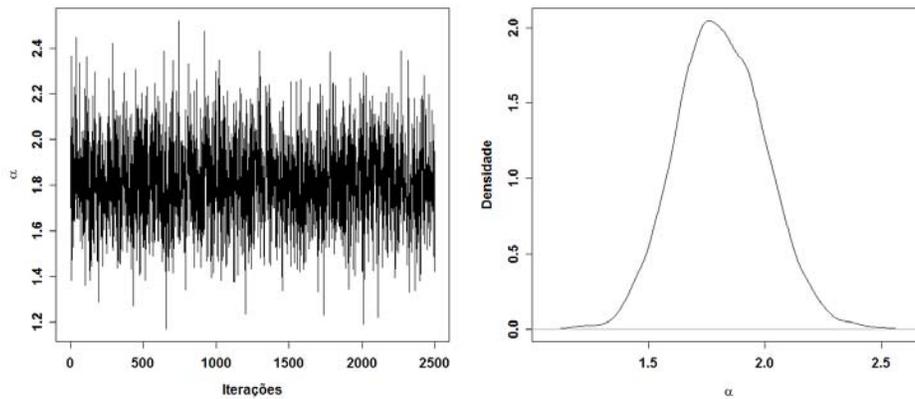


FIGURA 5 Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\alpha$ , que mede o efeito da acumulação de eventos ocorridos.

Os gráficos do traço da cadeia e da densidade para o parâmetro  $\delta$  são apresentados na FIGURA 6. Os gráficos do traço da cadeia e da densidade para o parâmetro  $\eta$  são apresentados na FIGURA 7.

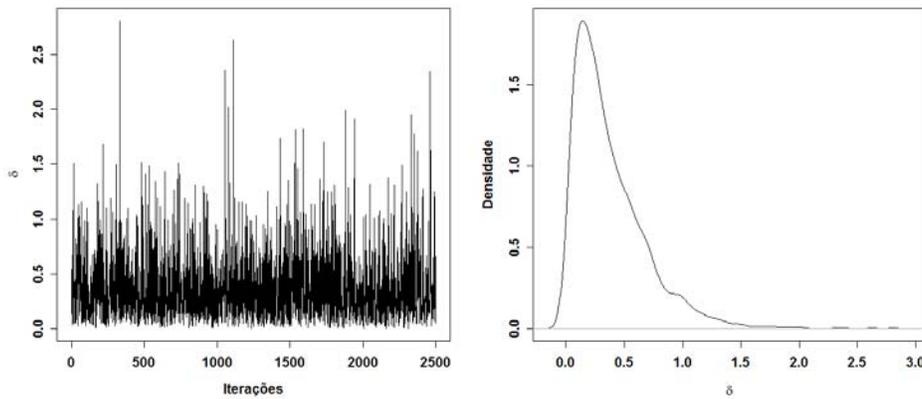


FIGURA 6 Traço da cadeia gerada pelo método MCMC e densidade *a posteriori* para o parâmetro  $\delta$  da distribuição Weibull.

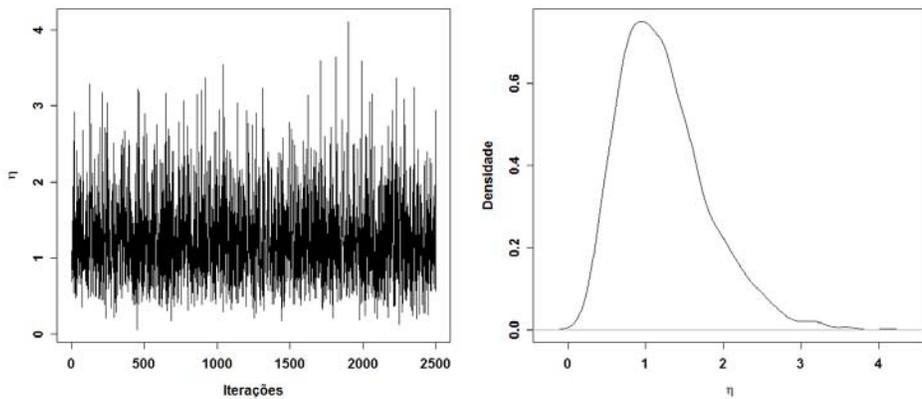


FIGURA 7 Traço da cadeia gerada pelo método MCMC e densidade *a posteriori* para o parâmetro  $\eta$ , proveniente da distribuição de fragilidade.

Observa-se que as cadeias geradas para todos os parâmetros do modelo geral para eventos recorrentes oscilaram em torno das médias estimadas, não apresentando picos discrepantes ao longo das cadeias.

Na TABELA 3 apresentam-se os valores da média a *posteriori* e intervalos de credibilidade a 95% para as fragilidades  $\alpha_i$  referentes ao modelo geral para eventos recorrentes.

TABELA 3 Média a *posteriori* e intervalos de credibilidade a 95%, para as fragilidades  $\alpha_i$ .

Indivíduo	Média	Intervalo de credibilidade	Indivíduo	Média	Intervalo de credibilidade
1	0,4959	[0,1374; 0,7698]	76	1,2694	[0,3909; 2,0505]
2	0,8	[0,6008; 0,9339]	77	0,9551	[0,0921; 1,8268]
3	0,724	[0,562; 0,8615]	78	0,9551	[0,0921; 1,8268]
4	0,5384	[0,2182; 0,7735]	79	1,4384	[1,1118; 1,9835]
5	0,7807	[0,5594; 0,9219]	80	1,4384	[1,1118; 1,9835]
6	0,1904	[0,0063; 0,4488]	81	1,4384	[1,1118; 1,9835]
7	0,7769	[0,5931; 0,914]	82	1,7241	[1,2832; 2,3557]
8	0,2534	[0,0153; 0,5134]	83	1,6649	[1,2645; 2,168]
9	0,3396	[0,0891; 0,5381]	84	1,6649	[1,2645; 2,168]
10	0,1687	[0,0045; 0,3922]	85	0,9188	[0,3301; 1,279]
11	0,7766	[0,6076; 0,9025]	86	0,246	[0,0087; 0,543]
12	0,8744	[0,794; 0,9451]	87	0,2928	[0,0175; 0,6451]
13	0,2619	[0,0401; 0,4752]	88	0,3116	[0,0178; 0,6651]
14	0,699	[0,4302; 0,8985]	89	0,3116	[0,0178; 0,6651]
15	0,8092	[0,6166; 0,9356]	90	1,8961	[1,4401; 2,3971]
16	0,2805	[0,0163; 0,5834]	91	1,5362	[1,2122; 1,9282]
17	0,8375	[0,704; 0,9364]	92	1,6549	[1,0843; 2,2919]

Indivíduo	Média	Intervalo de credibilidade	Indivíduo	Média	Intervalo de credibilidade
18	0,3293	[0,0269; 0,6092]	93	0,8103	[0,6645; 0,9187]
19	0,9269	[0,8413; 0,9802]	94	1,8961	[1,4401; 2,3971]
20	0,4625	[0,1811; 0,6609]	95	0,905	[0,5041; 1,1407]
21	0,3994	[0,1176; 0,6742]	96	0,7439	[0,3698; 0,9346]
22	0,3822	[0,0425; 0,6696]	97	0,1283	[0,0011; 0,36]
23	0,1493	[0,0038; 0,3571]	98	1,5709	[0,7093; 1,9863]
24	0,8825	[0,5589; 1,0835]	99	1,6931	[1,2417; 2,3853]
25	0,5606	[0,279; 0,768]	100	2,2818	[1,6345; 3,219]
26	0,8615	[0,6988; 0,9602]	101	1,6388	[1,0481; 2,2807]
27	0,252	[0,0153; 0,5572]	102	0,855	[0,3128; 1,3475]
28	0,624	[0,3647; 0,833]	103	2,0439	[1,4699; 2,8154]
29	0,624	[0,3647; 0,833]	104	2,1889	[1,5204; 3,2209]
30	0,8506	[0,6904; 0,9486]	105	0,8531	[0,7555; 0,9385]
31	0,6236	[0,3264; 0,8721]	106	2,5795	[1,8652; 3,4149]
32	0,1071	[0,0054; 0,246]	107	0,8863	[0,7929; 0,9496]
33	2,7791	[1,8905; 3,9306]	108	1,7933	[1,3289; 2,4022]
34	3,0086	[1,9834; 4,2292]	109	1,7933	[1,3289; 2,4022]
35	0,6815	[0,3229; 0,8767]	110	0,8863	[0,7929; 0,9496]
36	1,3157	[0,6626; 1,6783]	111	0,1104	[0,006; 0,2367]
37	0,8469	[0,4498; 1,1529]	112	0,4363	[0,183; 0,6769]
38	0,7757	[0,3506; 1,1615]	113	0,4971	[0,1508; 0,8072]
39	0,4482	[0,1186; 0,7697]	114	0,4876	[0,0661; 0,9103]
40	0,1636	[0,0055; 0,3692]	115	0,2556	[0,0124; 0,5497]
41	0,2271	[0,0106; 0,4775]	116	1,3554	[0,5641; 2,0412]
42	0,2916	[0,0224; 0,5646]	117	1,7254	[1,2912; 2,2921]
43	0,3805	[0,0457; 0,6962]	118	1,7254	[1,2912; 2,2921]
44	0,2916	[0,0224; 0,5646]	119	2,6986	[1,8351; 3,8652]
45	0,3805	[0,0457; 0,6962]	120	0,358	[0,0408; 0,7272]

Indivíduo	Média	Intervalo de credibilidade	Indivíduo	Média	Intervalo de credibilidade
46	1,0839	[0,5819; 1,5634]	121	1,8297	[1,3032; 2,5647]
47	0,8282	[0,6837; 0,9291]	122	0,2495	[0,0225; 0,5105]
48	2,0398	[1,5035; 2,5634]	123	0,3125	[0,02; 0,5955]
49	1,7983	[1,3206; 2,3764]	124	0,1791	[0,0123; 0,3843]
50	0,8266	[0,6461; 0,9402]	125	0,4149	[0,0402; 0,7537]
51	1,9717	[1,5667; 2,372]	126	0,5812	[0,0833; 0,9232]
52	1,0464	[0,4918; 1,6154]	127	0,5051	[0,0363; 0,9201]
53	1,8021	[1,3256; 2,4358]	128	0,3805	[0,1082; 0,6314]
54	0,8301	[0,7324; 0,9162]	129	0,6823	[0,3289; 0,9595]
55	0,8301	[0,7324; 0,9162]	130	0,2641	[0,0502; 0,4721]
56	0,896	[0,8307; 0,9512]	131	0,7815	[0,4749; 1,003]
57	1,846	[1,3021; 2,6659]	132	0,5125	[0,2312; 0,7326]
58	2,5038	[1,6067; 3,8862]	133	1,5974	[1,1902; 2,1735]
59	2,5038	[1,6067; 3,8862]	134	1,5974	[1,1902; 2,1735]
60	2,7603	[1,7529; 4,3706]	135	1,2494	[0,8996; 1,6572]
61	1,846	[1,3021; 2,6659]	136	1,6338	[1,245; 2,1269]
62	0,3346	[0,0556; 0,5975]	137	0,0253	[0,005; 0,0695]
63	0,5252	[0,1661; 0,8109]	138	1,6987	[1,2474; 2,3002]
64	0,0763	[0,0026; 0,1836]	139	0,2264	[0,0339; 0,4428]
65	0,3346	[0,0556; 0,5975]	140	1,3668	[1,0014; 1,8113]
66	0,5375	[0,1889; 0,7905]	141	0,7681	[0,6268; 0,8885]
67	0,5375	[0,1889; 0,7905]	142	1,6987	[1,2474; 2,3002]
68	0,5451	[0,1619; 0,8695]	143	1,423	[1,0965; 1,959]
69	0,4443	[0,1207; 0,6807]	144	1,423	[1,0965; 1,959]
70	0,6435	[0,2695; 0,9168]	145	1,4267	[0,9113; 2,1487]
71	0,4199	[0,1003; 0,7382]	146	1,4523	[1,0526; 2,1662]
72	0,1444	[0,0089; 0,318]	147	0,773	[0,6204; 0,9074]
73	1,413	[0,7698; 1,8758]	148	1,084	[0,6048; 1,6205]

Indivíduo	Média	Intervalo de credibilidade	Indivíduo	Média	Intervalo de credibilidade
74	0,3104	[0,1065; 0,4889]	149	0,773	[0,6204; 0,9074]
75	0,3381	[0,1283; 0,5537]			

Por meio dos intervalos de credibilidade apresentados na TABELA 3 pode-se observar que, para todos os indivíduos em tratamento, a variável de fragilidade foi estatisticamente significativa. O maior valor de fragilidade foi observado para o indivíduo 34, sendo  $Z_{34} = 3,0086$ . O menor valor, por outro lado, foi observado para o indivíduo 137, com  $Z_{137} = 0,0253$ .

Dos 149 pacientes, 52 apresentaram valores da variável de fragilidade maiores que 1 e, portanto, tendem a experimentar eventos mais rapidamente. Para os 97 pacientes que apresentaram valores de fragilidade menores do que 1, o tempo até a ocorrência dos eventos é mais longo.

Os pacientes 42 e 44 apresentaram valores de fragilidades iguais ( $Z_{42} = Z_{44} = 0,2916$ ), apesar de não compartilharem as mesmas covariáveis. Por exemplo, o paciente 42 é do sexo feminino, casado, branco, tem 60 anos, reside na cidade de Lavras e é hipertenso, enquanto o paciente 44 também é do sexo feminino e da raça branca, mas é solteiro, tem apenas 20 anos e foi levado ao tratamento de hemodiálise por outra doença, diferente da hipertensão.

#### 4.2 Modelo de fragilidade multiplicativo

Na TABELA 4 são apresentadas as estimativas da média, desvio padrão, intervalo de credibilidade de 95% (HPD), valores de  $\hat{R}$  para a verificação da convergência da cadeia por meio do critério de Gelman e Rubin e taxa de aceitação para os parâmetros do modelo.

Os gráficos do traço da cadeia e da densidade para os parâmetros  $\beta_1$  referente à covariável sexo e  $\beta_2$ , referente à covariável estado civil, são

apresentados nas FIGURAS de 8 e 9. Os gráficos relacionados aos demais parâmetros do vetor  $\beta$  são apresentados no Anexo B.

TABELA 4 Estimativas da média, desvio padrão (DP), intervalo de credibilidade 95% (IC), teste de diagnóstico (R) e taxa de aceitação (TA), para os parâmetros do modelo.

Parâmetro	Média	DP	IC	R	TA
$\beta_0$	-1,23	0,14	[-1,42; -0,88]	1,01	73,8
$\beta_1$	-0,80	0,10	[-1,02; -0,62]	1,01	73,8
$\beta_2$	2,37	0,70	[1,53; 3,69]	1,01	73,8
$\beta_3$	1,54	0,53	[0,74; 2,21]	1,01	73,8
$\beta_4$	0,03	0,01	[0,02; 0,05]	1,01	73,8
$\beta_5$	0,01	0,01	[-0,01; 0,01]	1,01	73,8
$\beta_6$	0,21	0,17	[0,05; 0,62]	1,01	73,8
$\beta_7$	-2,10	0,64	[-3,14; -1,19]	1,01	73,8
$\beta_8$	-1,93	0,18	[-2,30; -1,65]	1,01	73,8
$\beta_9$	-2,59	0,19	[-2,86; -2,21]	1,01	73,8
$\beta_{10}$	-2,12	0,59	[-3,33; -1,40]	1,01	73,8
$\sigma$	0,37	0,01	[0,33; 0,41]	0,99	47,82
$\eta$	3,27	0,99	[1,81; 5,15]	1,01	73,38

Com base nos resultados observados na TABELA 4, pode-se concluir que as covariáveis que estão associadas com um elevado risco de reinternação são estado civil ( $\beta_2$ ), raça ( $\beta_3$ ), idade ( $\beta_4$ ) e doença nefrite ( $\beta_6$ ), pois os valores para as respectivas médias são positivos. Como se pode observar, a única covariável que não é estatisticamente significativa é a procedência.

O risco de reinternações de pacientes do sexo masculino é menor do que o risco de pacientes femininos. Pacientes solteiros/viúvos/separados têm 10

vezes mais risco de reinternações do que pacientes casados. Pacientes negros têm risco de reinternação 4,66 vezes maior que brancos/pardos. Pacientes com nefrite têm risco de reinternação 1,23 vez maior do que os pacientes que não possuem tal doença.

O valor estimado para a variância da distribuição de fragilidade, dado por  $\eta = 3,27$ , e o intervalo de credibilidade [1,81;5,15] indicam que existe uma heterogeneidade não observada significativa entre os indivíduos.

A convergência das cadeias de todos os parâmetros do modelo foi verificada por meio do critério de Gelman & Rubin que, em todos os casos, apresentou valores de  $\hat{R}$  próximos a 1 e também por meio da visualização gráfica do traço apresentados a seguir.

Os valores das taxas de aceitação apresentaram resultados satisfatórios, como se observa na TABELA 4.

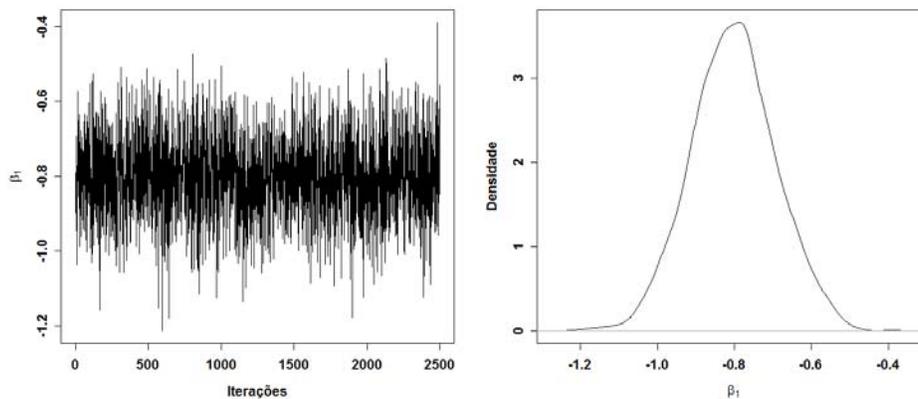


FIGURA 8 Cadeia gerada pelo método MCMC e densidade *a posteriori* para o parâmetro  $\beta_1$ , referente à covariável sexo.

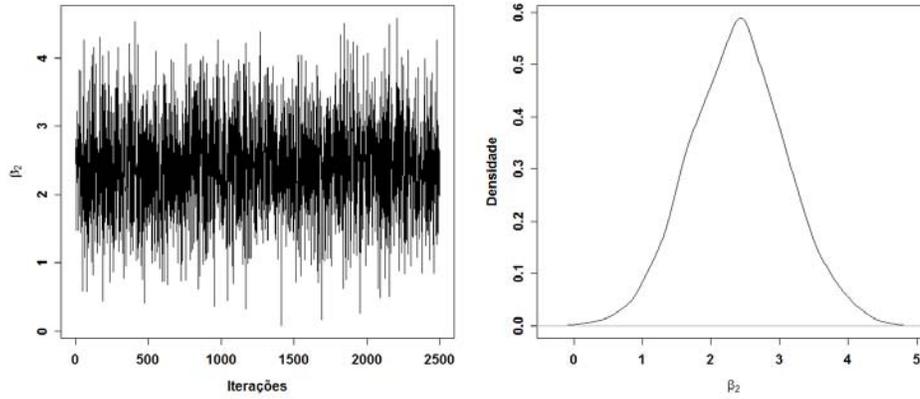


FIGURA 9 Cadeia gerada pelo método MCMC e densidade *a posteriori* para o parâmetro  $\beta_2$ , referente à covariável estado civil.

Os gráficos do traço da cadeia e da densidade para o parâmetro  $\delta$  são apresentados na FIGURA 10.

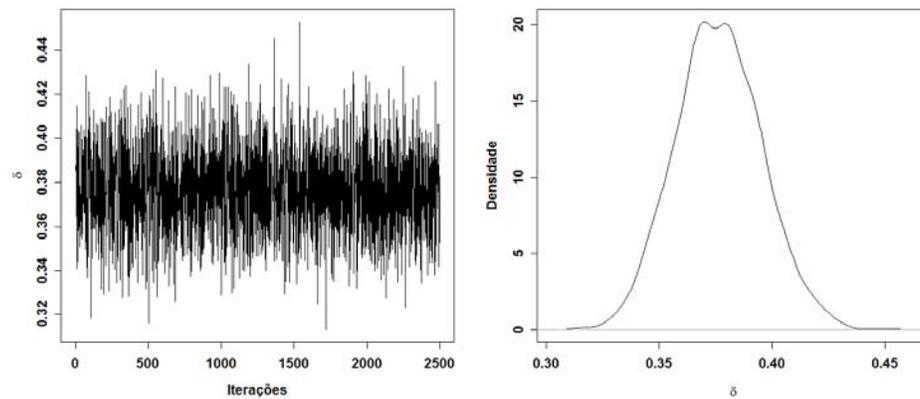


FIGURA 10 Cadeia gerada pelo método MCMC e densidade *a posteriori* para o parâmetro  $\delta$  da distribuição Weibull.

Os gráficos do traço da cadeia e da densidade para o parâmetro  $\eta$  são apresentados na FIGURA 11.

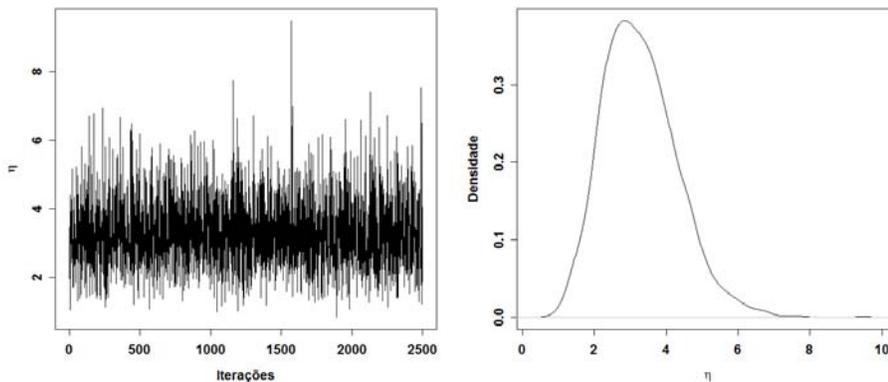


FIGURA 11 Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\eta$ , proveniente da distribuição de fragilidade.

Observa-se que as cadeias geradas para todos os parâmetros do modelo de fragilidade multiplicativo oscilaram em torno das médias estimadas, não apresentando picos discrepantes ao longo das cadeias.

Na TABELA 5, apresentam-se os valores da média a *posteriori* e intervalos de credibilidade, a 95%, para as fragilidades  $z_i^*$ , referentes ao modelo de fragilidade multiplicativo.

TABELA 5 Média a *posteriori* e intervalos de credibilidade, a 95%, para as fragilidades  $z_i^*$ .

Indivíduo	Média	Intervalo de credibilidade	Indivíduo	Média	Intervalo de credibilidade
1	0,0224	[0,0022; 0,0455]	76	0,6787	[0,1622; 1,6682]
2	0,131	[0,0132; 0,29]	77	0,5544	[0,0182; 1,1029]

Indivíduo	Média	Intervalo de credibilidade	Indivíduo	Média	Intervalo de credibilidade
3	0,1381	[0,07; 0,2111]	78	0,5544	[0,0182; 1,1029]
4	0,0412	[0,0053; 0,0849]	79	0,5432	[0,3238; 0,7792]
5	0,1175	[0,0099; 0,2644]	80	0,5432	[0,3238; 0,7792]
6	0,0041	[0; 0,0136]	81	0,5432	[0,3238; 0,7792]
7	0,1737	[0,029; 0,3117]	82	0,9554	[0,5561; 1,4077]
8	0,0102	[0,004; 0,0279]	83	1,3676	[0,826; 2,0773]
9	0,0295	[0,0033; 0,0634]	84	1,3676	[0,826; 2,0773]
10	0,0039	[0; 0,0123]	85	0,1979	[0,0074; 0,4047]
11	0,2034	[0,0534; 0,3338]	86	0,0063	[0; 0,019]
12	0,2236	[0,133; 0,314]	87	0,0368	[0,003; 0,1038]
13	0,0118	[0,005; 0,0328]	88	0,0215	[0,002; 0,0616]
14	0,077	[0,0045; 0,1869]	89	0,0215	[0,002; 0,0616]
15	0,1382	[0,0167; 0,305]	90	2,3725	[1,2479; 3,2966]
16	0,004	[0,001; 0,011]	91	0,9537	[0,4656; 1,3261]
17	0,0975	[0,0305; 0,1936]	92	3,0819	[1,646; 4,2922]
18	0,0089	[0,003; 0,0201]	93	0,2443	[0,0749; 0,3823]
19	0,3339	[0,0921; 0,5334]	94	2,3725	[1,2479; 3,2966]
20	0,0543	[0,0095; 0,1078]	95	0,4439	[0,1271; 0,771]
21	0,0201	[0,0015; 0,0451]	96	0,1282	[0,0065; 0,2656]
22	0,0226	[0,001; 0,0588]	97	3,8717	[0,4776; 6,886]
23	0,0032	[0; 0,0102]	98	0,8088	[0,0822; 1,4461]
24	0,2645	[0,0786; 0,4719]	99	0,7842	[0,4522; 1,1592]
25	0,0707	[0,0122; 0,1464]	100	1,9766	[1,1317; 3,0416]
26	0,2129	[0,0801; 0,3615]	101	3,169	[1,7842; 4,9236]
27	0,0157	[0,001; 0,0497]	102	4,3613	[2,4493; 6,8155]
28	0,0335	[0,006; 0,0718]	103	2,5728	[1,4587; 3,9841]
29	0,0335	[0,006; 0,0718]	104	1,3804	[0,8011; 2,1059]
30	0,1792	[0,0266; 0,3739]	105	0,1881	[0,106; 0,2727]

Indivíduo	Média	Intervalo de credibilidade	Indivíduo	Média	Intervalo de credibilidade
31	0,1436	[0,0244; 0,2815]	106	3,3598	[2,0223; 4,363]
32	0,0091	[1e-04; 0,0237]	107	0,3406	[0,1236; 0,5108]
33	2,5466	[1,3466; 3,3928]	108	1,347	[0,7769; 1,7693]
34	5,8094	[3,1895; 7,7694]	109	1,347	[0,7769; 1,7693]
35	0,0597	[0,0112; 0,1096]	110	0,3406	[0,1236; 0,5108]
36	0,3968	[0,1398; 0,6008]	111	0,0371	[0,007; 0,0904]
37	0,156	[0,0174; 0,3476]	112	0,3484	[0,0993; 0,6117]
38	0,3802	[0,0472; 0,8331]	113	1,1453	[0,3787; 1,9066]
39	0,6044	[0,0753; 1,318]	114	0,4096	[0,018; 0,8022]
40	0,1782	[0,0063; 0,4235]	115	0,5964	[0,0265; 1,1642]
41	0,1452	[0,0051; 0,3459]	116	5,495	[3,0873; 7,5083]
42	0,0131	[0,002; 0,0353]	117	1,163	[0,5704; 1,5656]
43	0,0461	[0,0014; 0,113]	118	1,163	[0,5704; 1,5656]
44	0,0131	[0,002; 0,0353]	119	2,816	[1,6587; 4,5526]
45	0,0461	[0,0014; 0,113]	120	8,771	[4,8913; 14,2883]
46	2,5531	[0,8594; 3,9702]	121	1,1146	[0,6833; 1,7182]
47	0,2322	[0,0528; 0,3866]	122	0,1724	[0,0033; 0,3731]
48	2,233	[0,949; 3,1985]	123	0,0154	[0,001; 0,0375]
49	2,8999	[1,2535; 4,1613]	124	0,2117	[0,0042; 0,457]
50	0,1536	[0,0186; 0,3301]	125	0,0547	[0,009; 0,1206]
51	0,9682	[0,2256; 1,8204]	126	0,0938	[0,0052; 0,2169]
52	3,8197	[2,0909; 5,97]	127	0,2285	[0,014; 0,5168]
53	2,6095	[1,4364; 4,0505]	128	0,8305	[0,2215; 1,4712]
54	0,189	[0,1189; 0,2578]	129	0,5211	[0,1364; 0,9304]
55	0,189	[0,1189; 0,2578]	130	0,9852	[0,2634; 1,7357]
56	0,3534	[0,2428; 0,4609]	131	0,2116	[0,051; 0,3873]
57	1,4966	[0,9418; 2,3116]	132	0,0569	[0,0069; 0,1199]
58	2,6398	[1,5533; 4,1883]	133	0,81	[0,4471; 1,1581]

Indivíduo	Média	Intervalo de credibilidade	Indivíduo	Média	Intervalo de credibilidade
59	2,6398	[1,5533; 4,1883]	134	0,81	[0,4471; 1,1581]
60	3,783	[2,2483; 6,1445]	135	1,5082	[0,8687; 2,2376]
61	1,4966	[0,9418; 2,3116]	136	1,3234	[0,7635; 1,9696]
62	0,2649	[0,0397; 0,623]	137	5,7103	[3,2145; 8,6666]
63	0,1444	[0,021; 0,3442]	138	0,9246	[0,5384; 1,3571]
64	0,5059	[0,0802; 1,1893]	139	3,7162	[2,0995; 5,6192]
65	0,2649	[0,0397; 0,623]	140	1,7222	[1,0213; 2,6114]
66	0,0841	[0,0112; 0,2029]	141	0,1269	[0,0714; 0,1888]
67	0,0841	[0,0112; 0,2029]	142	0,9246	[0,5384; 1,3571]
68	0,2319	[0,0278; 0,5827]	143	0,5257	[0,3143; 0,7527]
69	0,0272	[0,0015; 0,0754]	144	0,5257	[0,3143; 0,7527]
70	0,0954	[0,0099; 0,2437]	145	2,0227	[1,1756; 3,0655]
71	0,3001	[0,0349; 0,7509]	146	0,6191	[0,3634; 0,9038]
72	0,5049	[0,0599; 1,2565]	147	0,1513	[0,0708; 0,2457]
73	1,273	[0,2312; 2,5704]	148	2,4905	[1,4477; 3,7914]
74	0,1185	[0,0146; 0,2272]	149	0,1513	[0,0708; 0,2457]
75	0,0686	[0,01; 0,1368]			

Por meio dos intervalos de credibilidade apresentados na TABELA 5 pode-se observar que, para todos os indivíduos em tratamento, a variável de fragilidade foi estatisticamente significativa. O maior valor de fragilidade foi observado para o indivíduo 120, sendo de  $Z_{120} = 8,771$ . O menor valor, por outro lado, foi observado para o indivíduo 23, com  $Z_{23} = 0,0032$ .

Dos 149 pacientes, 41 apresentaram valores da variável de fragilidade maiores que 1 e, portanto, tendem a experimentar eventos mais rapidamente. Para os 108 pacientes que apresentaram valores de fragilidade menores do que 1, o tempo até a ocorrência dos eventos é mais longo.

Os pacientes 117 e 118 apresentaram valores de fragilidades iguais ( $Z_{117} = Z_{118} = 1,163$ ), apesar de não compartilharem as mesmas covariáveis. Por exemplo, o paciente 117 é do sexo masculino, casado, branco, tem 65 anos, reside na microrregião de Lavras e foi levado ao tratamento de hemodiálise por causa da nefrite, enquanto o paciente 118 é do sexo feminino, tem raça branca, reside na microrregião de Lavras, mas é viúvo, tem 79 anos e tem diabetes.

### 4.3 Modelo de Regressão de Cox

Na TABELA 6 são apresentadas as estimativas da média, desvio padrão, intervalo de credibilidade, a 95% (HPD), valores de  $\hat{R}$  para a verificação da convergência da cadeia por meio do critério de Gelman e Rubin e taxa de aceitação para os parâmetros do modelo.

Os gráficos do traço da cadeia e da densidade para os parâmetros  $\beta_1$ , referente à covariável sexo e  $\beta_2$ , referente à covariável estado civil, são apresentados nas FIGURAS 11 e 12. Os gráficos relacionados aos demais parâmetros do vetor  $\beta$ , bem como os gráficos para  $\delta$ , são apresentados no Anexo C.

TABELA 6 Estimativas da média, desvio padrão (DP), intervalo de credibilidade, a 95% (IC), teste de diagnóstico (R) e taxa de aceitação (TA) para os parâmetros do modelo.

Parâmetro	Média	DP	IC	R	TA
$\beta_0$	-1,67	0,18	[-1,97; -1,38]	1,01	52,56
$\beta_1$	-1,05	0,11	[-1,31; -0,85]	1,01	52,56
$\beta_2$	5,52	0,18	[5,16; 5,81]	1,01	52,56
$\beta_3$	2,82	0,12	[2,58; 3,04]	1,01	52,56
$\beta_4$	3,92	0,16	[3,01; 4,56]	1,01	52,56
$\beta_5$	0,01	0,01	[-0,01; 0,03]	1,01	52,56

$\beta_6$	1,83	0,11	[1,60; 2,06]	1,01	52,56
$\beta_7$	-3,66	0,50	[-4,52; -2,88]	1,01	52,56
$\beta_8$	-3,10	0,17	[-3,53; -2,84]	1,01	52,56
$\beta_9$	-3,39	0,18	[-3,67; -2,84]	1,01	52,56
$\beta_{10}$	-3,89	0,17	[-4,23; -3,59]	1,01	52,56
$\delta$	0,32	0,59	[0,01; 1,91]	0,99	67,85

Com base nos resultados observados na TABELA 6, pode-se concluir que as covariáveis que estão associadas com um elevado risco de reinternação são estado civil ( $\beta_2$ ), raça ( $\beta_3$ ), idade ( $\beta_4$ ) e doença nefrite ( $\beta_6$ ). A única covariável que não é estatisticamente significativa é a procedência ( $\beta_5$ ), com intervalo de credibilidade [-0,01; 0,03].

Os valores das taxas de aceitação apresentaram resultados satisfatórios, como pode ser visto na TABELA 6.

A convergência das cadeias de todos os parâmetros do modelo foi verificada por meio do critério de Gelman & Rubin, em que em todos os casos apresentaram valores de  $\hat{R}$  próximos a 1 e também por meio da visualização gráfica do traço da cadeia apresentada a seguir.

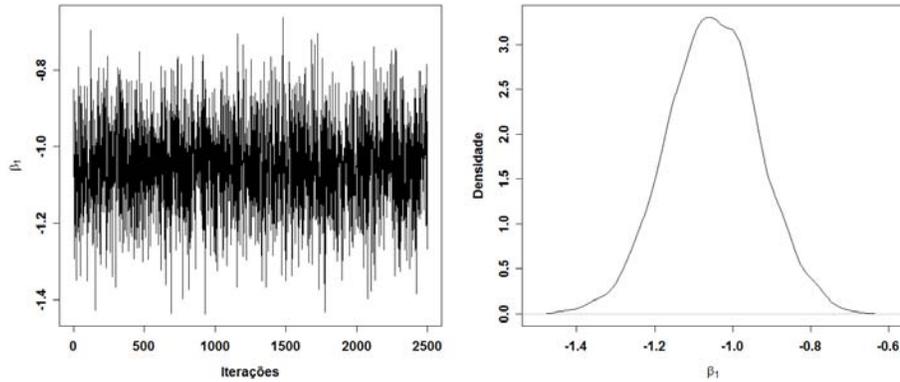


FIGURA 12 Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_1$ , referente à covariável sexo.

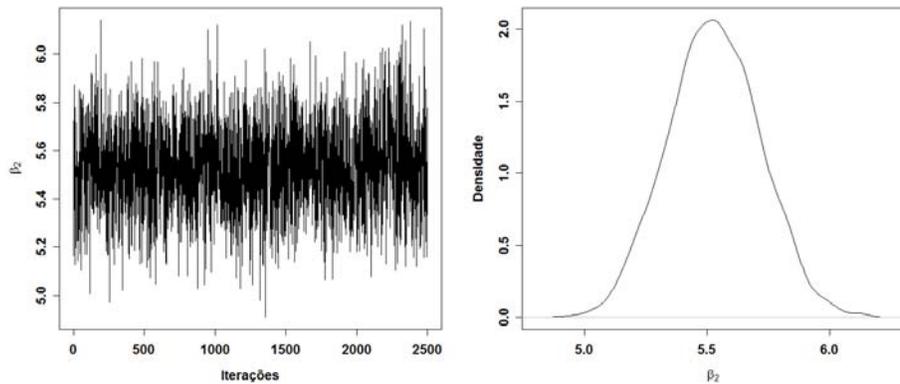


FIGURA 13 Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_2$ , referente à covariável estado civil.

Observa-se que as cadeias geradas para todos os parâmetros do modelo de Cox oscilaram em torno das médias estimadas, não apresentando picos discrepantes ao longo delas.

#### 4.4. Comparação de modelos

A escolha do melhor modelo foi feita com base no AIC e no BIC, os quais foram descritos na seção 2.8.6. Os resultados podem ser vistos na TABELA 7.

TABELA 7 Critério de informação de Akaike (AIC) e critério de informação bayesiano (BIC) para comparação do modelo geral para eventos recorrentes, o modelo de fragilidade mutiplicativo e o modelo de Cox.

Critério	Modelo geral	Modelo de fragilidade	Modelo de Cox
AIC	4401,96	4522,50	4800,85
BIC	4444,01	4561,557	4996,65

Como foram verificados menores valores de BIC e AIC, o modelo que melhor se adapta aos dados é o modelo geral para eventos recorrentes, proposto por Peña & Hollander (2004), o qual incorpora o parâmetro que mede o efeito do impacto da acumulação de tempos de ocorrências de eventos sobre os indivíduos.

Os maiores valores de BIC e AIC foram verificados para o modelo de regressão de Cox, o que pode ser explicado pelo fato de que esse modelo não considera que o tempo e a ocorrência de eventos não estão relacionados com a formação inicial e subsequentes ocorrências.

## 5 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foi estudada, sob uma abordagem bayesiana, a classe geral de modelos para eventos recorrentes com fragilidade multiplicativa, proposta por Peña & Hollander (2004). Essa classe de modelos inclui, como caso particular, o modelo de fragilidade multiplicativo, proposto por Clayton (1978) e o modelo de regressão de Cox (1972), também estudados nesta tese.

Concluiu-se que a classe geral de modelos é importante porque considera, além dos efeitos das covariáveis, o efeito da intervenção que é realizada após cada recorrência do evento de interesse, por meio da noção de idade efetiva e considera também o efeito da acumulação da recorrência dos eventos sobre os indivíduos, além da fragilidade não observada, que pode induzir correlações entre os tempos de eventos.

A aplicação ao conjunto de dados referente aos pacientes com insuficiência renal crônica, tratados no Centro de Terapia Renal Substitutiva da Santa Casa de Misericórdia de Lavras, em Lavras, MG, demonstrou a importância dos termos de fragilidade e do parâmetro que mede efeito da acumulação da recorrência dos eventos sobre os indivíduos, dados, respectivamente, por  $\eta = 1,23$ , com IC [0,50; 2,32] e  $\alpha = 1,80$  com IC [1,49; 2,19] no modelo geral.

Para os três modelos estudados verificou-se que as covariáveis que estão associadas com um elevado risco de reinternação são estado civil, raça, idade e doença nefrite. A única covariável que não foi considerada estatisticamente significativa foi procedência.

Foi mostrado também que pacientes podem apresentar valores de fragilidades individuais iguais, mesmo não compartilhando as mesmas

covariáveis. Esse resultado foi verificado tanto para a classe geral de modelos quanto para o modelo de fragilidade multiplicativo.

Em particular, concluiu-se que o modelo proposto por Peña & Hollander (2004) pode ser útil para analisar dados de tempos de recorrência de hospitalizações para pacientes em tratamento de hemodiálise. Nesse caso, o efeito do acúmulo de recorrências de evento é um fator importante para prever uma nova internação.

Foram utilizados os critérios AIC e BIC para comparação dos modelos, ou seja, classe geral de modelos para eventos recorrentes, modelo de fragilidade multiplicativo e modelo de regressão de Cox. Ambos os critérios selecionaram, para utilização nos dados de insuficiência renal, o modelo geral para eventos recorrentes proposto por Peña & Hollander (2004), o qual incorpora o parâmetro que mede o efeito do impacto da acumulação de tempos de ocorrências de eventos sobre os indivíduos.

Neste trabalho, foi utilizada a formulação da idade efetiva considerando reparação mínima. Entretanto, para esse conjunto de dados, poderia ser estudada a formulação da idade efetiva considerando reparação perfeita. Nesse sentido, o modelo torna-se mais complexo para a função de idade efetiva, na qual a resposta do tratamento após cada recorrência é incluída no modelo, atuando na função risco de base.

Uma importante proposta para trabalho futuro é considerar outra forma para a função risco de base e também para a variável de fragilidade, além de considerar outra abordagem para a construção de *prioris*, como, por exemplo, a *priori* de referencia (Bernardo, 1979) ou a *priori* de Jeffreys (1961).

Para um estudo futuro, seria interessante também considerar o tempo final de estudo como sendo aleatório para cada indivíduo e não fixo, para todos os indivíduos, como foi considerado neste trabalho.

## REFERÊNCIAS BIBLIOGRÁFICAS

AALEN, O. O. Heterogeneity in survival analysis. **Statistics in Medicine**, New York, v. 7, n. 11, p. 1121-1137, Nov. 1988.

AALEN, O. O. Model for nonparametric regression analysis of counting processes. **Lecture Notes in Statistics**, New York, v. 2, p. 1-25, Jan. 1980.

AALEN, O. O. Nonparametric inference for a family of counting processes. **Annals of Statistics**, Hayward, v. 6, n. 3, p. 701-726, July/Sept. 1978.

AALEN, O. O.; FOSEN, J.; WEEDON-FEKJAER, H.; BORGAN, O.; HUSEBYE, E. Dinamic analysis of multivariate failure time data. **Biometrics**, Washington, v. 60, n. 3, p. 764-773, July/Sept. 2004.

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, New York, v. 19, n. 6, p. 716-723, June 1974.

ALLISON, P. D. **Survival analysis using the SAS® sistem**: a practical guide. Cary: SAS Institute, 1995. 292 p.

ANDERSEN, P. K.; BORGAN, O.; GILL, R. D.; KEIDING, N. **Statistical models based on counting process**. New York: Springer-Verlag, 1993.

ANDERSEN, P. K.; GILL, R. D. Cox's regression model counting process: a large sample study. **Annals of Statistics**, Hayward, v. 10, n. 4, p. 1100-1120, Oct./Dec. 1982.

ANDRADE, P. C. de R. **Comparações múltiplas bayesianas em modelos normais homocedásticos e heterocedásticos**. 2008. 96 p. Tese (Doutorado em Estatística E Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras.

BARAI, U.; TEOH, N. Multiple statistics for multiple events, with application to repeated infectious in the growth factor studies. **Statistics in Medicine**, New York, v. 16, n. 8, p. 941-949, Apr. 1997.

BASTOS, L. S. **Modelos dinâmicos e estáticos de sobrevivência com fragilidade espacial**. 2003. 163 p. Dissertação (Mestrado em Ciências Estatísticas) – Universidade Federal do Rio de Janeiro, Rio de Janeiro.

BERGER, J. **Statistical decision theory and Bayesian analysis**. New York: Springer-Verlag, 1985. 609 p.

BERNARDO, J. M. Reference posterior distributions for Bayesian inference. **Journal of the Royal Statistical Society**, London, v. 41, n. 2, p. 113-147, Apr./June 1979.

BERNARDO, J. M.; SMITH, A. F. M. Bayesian theory. **Statistical Methods and Applications**, New York, v. 3, n.1, p. 155-160, Feb. 1994.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística**. Rio de Janeiro: Sociedade Brasileira de Matemática, 2001. 125 p.

BOX, G. E. P.; TIAO, G. C. **Bayesian inference in statistical analysis**. New York: J. Wiley, 1992. 603 p.

BOX-STEFFENSMEIER, J. M.; BOEF, S. de. A Monte Carlo analysis for recurrent events data. **The Society for Political Methodology**, Washington, July 2002. Disponível em: <URL:[http://http://polmeth.wustl.edu/retrieve.php?id=90](http://polmeth.wustl.edu/retrieve.php?id=90)>. Acesso em: 15 maio 2009.

BROEMELING, L. D. **Bayesian analysis of linear models**. New York: J. Wiley, 1989. 412 p.

BROWN, B. W.; FLOOD, M. M. Tumbler mortality. **Journal of the American Statistical Association**, New York, v. 42, n. 2, p. 562-574, June 1947.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel inference: understanding AIC and BIC in model selection. **Sociological Methods and Research**, Beverly Hills, v. 33, n. 2, p. 261-304, Feb. 2004.

CARDOZO, M. T.; VIEIRA, I. O.; CAMPANELLA, L. C. A. Alterações nutricionais em pacientes renais crônicos em programa de hemodiálise. **Revista Brasileira de Nutrição Clínica**, São Paulo, v. 21, n. 4, p. 284-289, abr. 2006.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de sobrevida: teoria e aplicações em saúde**. Rio de Janeiro: Fiocruz, 2005. 395 p.

CHIANG, C. T.; JAMES, L. F.; WANG, M. C. Random weighted bootstrap method for recurrent events with informative censoring. **Lifetime Data Analysis**, Amsterdam, v. 11, n. 4, p. 489-509, Dec. 2005.

- CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. **The American Statistician**, Alexandria, v. 49, n. 4, p. 327-335, Apr. 1995.
- CLAYTON, D. G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. **Biometrika**, London, v. 65, n. 1, p. 141-151, Apr. 1978.
- CLAYTON, D. G. A Monte Carlo method for Bayesian inference in frailty models. **Biometrics**, Washington, v. 47, n. 2, p. 467-485, Apr./June 1991.
- CLAYTON, D.; CUZICK, J. The em algorithm for cox's regression model using GLIM. **Journal Applied Statistics**, Abingdon, v. 34, n. 2, p. 148-156, 1985.
- COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: E. Blücher, 2006. 369 p.
- COOK, R. J.; LAWLESS, J. F. **The statistical analysis of recurrent events**. New York: Springer, 2007. 404 p.
- COX, D. R. Partial likelihood. **Biometrika**, London, v. 62, n. 1, p. 269-276, Apr. 1975.
- COX, D. R. Regression models and life tables. **Journal of the Royal Statistical Society**, London, v. 34, n. 1, p. 187-220, Jan./Mar. 1972.
- DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society**, London, v. 39, n. 1, p. 1-38, Jan./Mar. 1977.
- DUCHATEAU, L.; JANSSEN, P.; KEZIC, I.; FORTPIED, C. Evolution of recurrent asthma event rate over time in frailty models. **Journal Applied Statistics**, Abingdon, v. 52, n. 3, p. 355-363, Mar. 2003.
- DUCHESNE, T. Regression models for reliability given the usage accumulation history. In: WILSON, A.; MCNULTY, S. K.; ARMIJO, Y.; LIMNIOS, N. (Ed.). **Modern statistical and mathematical methods in reliability**. Singapore: World Scientific, 2005. 409 p.
- DUNSON, D. B.; CHEN, Z. Selecting factors predictive of heterogeneity in multivariate event time data. **Biometrics**, Washington, v. 60, n. 1, p. 352-358, Jan./Mar. 2004.

EHLERS, R. S. **Introdução à inferência bayesiana**. Paraná: Departamento de Estatística, 2005. Disponível em: <<http://leg.ufpr.br/~ehlers/bayes>>. Acesso em: 3 jun. 2009.

ELBERS, C.; RIDDER, G. True and spurious duration dependence: the identifiability of the proportional hazard model. **Review of Economic Studies**, Bristol, v. 49, n. 156, p. 403-409. Apr./June 1982.

FLEMING, T. R.; HARRINGTON, D. P. **Counting processes and survival analysis**. New York: J. Wiley, 1991. 429 p.

FOGO, J. C. **Modelo de regressão para um processo de renovação Weibull com termo de fragilidade**. 2007. 109 p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

FREDETTE, M.; LAWLESS, J. F. Finite-horizon prediction of recurrent events, with application to forecasts of warranty claims. **Technometrics**, Washington, v. 49, n. 1, p. 66-80, Jan./Mar. 2007.

GAMERMAN, D. **Markov Chain Monte Carlo: stochastic simulation for bayesian inference**. London: Chapman, 1997. 245 p.

GELFAND, A. E. Gibbs sampling. **Journal of the American Statistical Association**, New York, v. 95, n. 452, p. 1300-1304, Oct./Dec. 2000.

GELFAND, A. E.; SMITH A. F. M. Sampling-based approaches to calculating marginal densities. **Journal of the American Statistical Association**, New York,, v. 85, n. 410, p. 348-409, June/Sept. 1990.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. **Bayesian data analysis**. London: Chapman, 2000. 526 p.

GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distribution and the Bayesian restoration of images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, New York, v. 6, n. 6, p. 721-741, June 1984.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, Hayward, v. 7, n. 4, p. 457-511, Apr. 1992.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: BERNARDO, J. M.; BERGER, J. O.; DAWID, A. P.; SMITH, A. F. M. (Ed.). **Bayesian statistics**. New York: Oxford University, 1992. p. 625-631.

GHOSH, D.; LIN, D, Y. Marginal regression models for recurrent and terminal events. **Statistica Sinica**, Taipei, v. 12, n. 3, p. 663-688, July/Sept. 2002.

GIOLO, S. R. **Variáveis latentes em análise de sobrevivência e curvas de crescimento**. 2003. 100 p. Tese (Doutorado em Estatística e Experimentação Agrônômica) - Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba.

GONZALÉZ, J. R. **Inference for a general class of models for recurrent events with application to cancer data**. 2006. 203 p. Thesis (PhD) - Universitat Politècnica de Catalunya, Barcelona.

GONZALÉZ, J. R.; PEÑA, E. A.; SLATE, E. H. Modelling intervention effects after câncer relaps. **Statistics in Medicine**, New York, v. 24, n. 24, p. 3959-3975, Dec. 2005.

GOUVÊA, G. D. R. **Estimador *Bootstrap* não-paramétrico de curvas de sobrevivência para dados entomológicos com censura intervalar**. 2006. 61 p. Dissertação (Mestrado Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras.

GOUVÊA, G. D. R.; OLIVEIRA, F. L. P. de; VIVANCO, M. J. F. Análise de eventos competitivos: uma aplicação aos dados de hemodiálise da cidade de Lavras-MG. **Revista Brasileira de Biometria**, São Paulo, v. 27, n. 3, p. 491-500, mar. 2009.

HARRIS, E. K.; ALBERT, A. **Survivorship analysis for clinical studies**. New York: Marcel Dekker Incorporation, 1991. 197 p.

HASTINGS, W. K. Monte Carlo sampling methods using Markov chains and their applications. **Biometrika**, London, v. 57, n. 1, p. 97-109, Apr. 1970.

HEIDELBERG, P.; WELCH, P. Simulation run length control in the presence of an initial transient. **Operations Research**, Baltimore, v. 31, n. 6, p. 1109 -1144, Dec. 1993.

- HOOVER, D. R. Extension of the life table to repeating and changing events. **American Journal of Epidemiology**, Baltimore, v. 143, n. 12, p. 1266-1276, Dec. 1996.
- HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression**. New York: J. Wiley, 1989.
- HOUGAARD, P. A. **Analysis of multivariate survival data**. New York: Springer-Verlag; 2000.
- HOUGAARD, P. A. A class of multivariate failure time distributions. **Biometrika**, London, v. 73, n. 3, p. 671-678, Dec. 1986b.
- HOUGAARD, P. A. Life table methods for heterogeneous populations: distributions describing heterogeneity. **Biometrika**, London, v. 71, n. 1, p. 75-83, Apr. 1984.
- HOUGAARD, P. A. Survival models for heterogeneous populations derived from stable distributions. **Biometrika**, London, v. 73, n. 2, p. 387-396, Aug. 1986a.
- HOUGAARD, P. A.; HARVALD, B.; HOLM, N. V. Measuring the similarities between the lifetimes of adult Danish twins born between 1881-1930. **Journal of the American Statistical Association**, New York, v. 87, n. 417, p. 17-24, Jan./Mar. 1992.
- HUANG, C. Y.; WANG, M. C. Joint modeling and estimation for recurrent event processes and failure time data. **Journal of the American Statistical Association**, New York, v. 99, n. 464, p. 1153-1165, Dec. 2004.
- HUANG, X.; LIU, L. a joint frailty model for survival and gap times between recurrent events. **Biometrics**, Washington, v. 63, n. 1, p. 389-397, Jan./Mar. 2007.
- HUANG, X.; WOLFE, R. A. A frailty model for informative censoring. **Biometrics**, Washington, v. 58, n. 2, p. 510-520, Apr./June 2002.
- JACKSON, L. A.; YU, O.; HECKBERT, S. R.; PSATY, B. M.; MALAIS, D.; BARLOW, W. E.; THOMPSON, W. W. Influenza vaccination is not associated with a reduction in the risk of recurrent coronary events. **American Journal of Epidemiology**, Baltimore, v. 156, n. 7, p. 634-640, July 2002.

JACOD, J. Multivariate point processes: predictable projection, Radon-Nikodym derivatives, representation of martingales. **Probability Theory and Related Fields**, New York, v. 31, n. 3, p. 235-253, Sept. 1975.

JEFFREYS, H. **Theory of probability**. 3. ed. Oxford: Clarendon, 1961. 447 p.

JENSEN, H.; BROOKMEYER, R.; AABY, P.; ANDERSEN, P. K. Shared frailty models for left-truncated multivariate survival data. **Research Report**, Baltimore, v. 4, n. 2, p. 1-29, Apr. 2004.

JONES, G.; ROCKE, M. D. Multivariate survival analysis with double-censored data: application to the assesment of Accutane treatment for fibrodysplasia ossificans progressiva. **Statistics in Medicine**, New York, v. 21, n. 17, p. 2547-2562, Sept. 2002.

KALBFLEISCH, J. D.; PRENTICE, R. L. **The statistical analysis of failure time data**. 2. ed. New York: J. Wiley; 2002. 437 p.

KASS, R. E.; RAFTERY, A. E. Bayes factors and model uncertainty. **Journal of the American Statistical Association**, New York,, v. 90, n. 430, p. 773-795, Sept. 1995.

KELLY, P. J.; LIM, L. L. Y. Survival analysis for recurrent event data: an application to childhood infections diseases. **Statistics in Medicine**, New York, v. 19, n. 1, p.13-33, Jan. 2000.

KLEIN, J. P. Semiparametric estimation of random effects using the cox model based on the EM algorithm. **Biometrics**, Washington, v. 48, n. 3, p. 795-806, July/Sept. 1992.

KLEIN, J. P.; MOESCHBERGER, M. L. **Survival analysis: techniques for censored and truncated data**. New York: Springer, 1997. 118 p.

KIM, Y. J. Regression analysis of double censored failure time data with frailty. **Biometrics**, Washington, v. 62, n. 3, p. 458-464, July/Sept. 2006.

KOOPERBERG, C. Bivariate density estimation with an application to survival analysis. **Journal of Computational and Graphical Statistics**, Alexandria, v. 7, n. 1, p. 322-341, Jan./Mar. 1998.

LAWLESS, J. F. **Statistical models and methods for lifetime data**. New York: Wiley, 1982. 579.

LAWLESS, J. F.; NADEAUS, C. Some simple robust methods for the analysis of recurrent events. **Technometrics**, Washington, v. 37, n. 1, p. 158-168, Jan./Mar. 1995.

LEE, T. E.; WANG, W. J. **Statistical methods for survival data analysis**. 3. ed. New York: Wiley-Interscience; 2003. 534 p.

LIU, L.; WOLFE, R. A.; HUANG, X. Shared frailty models for recurrent events and a terminal event. **Biometrics**, Washington, v. 60, n. 3, p. 747-756, July/Sept. 2004.

MAHÉ, C.; CHEVRET, S. Estimating regression parameters and degree of dependence for multivariate failure time data. **Biometrics**, Washington, v. 55, n. 3, p. 1078-1084, July/Sept. 1999.

MANTON, K. G.; STALLARD, E.; VAUPEL, J. W. Alternative models for heterogeneity of counting process. **Annals of Statistics**, Hayward, v. 23, n. 4, p. 1712-1734, Oct./Dec. 1986.

MATUDA, N. S. **Fragilidade gama e variância robusta: extensões do modelo semiparamétrico de Cox**. 2005. 99 p. Tese (Doutorado em Saúde Pública) - Universidade de São Paulo, São Paulo.

MCGILCHRIST, C. A.; AISBETT, C. W. Regression with frailty in survival analysis. **Biometrics**, Washington, v. 47, n. 2, p. 461-466, Apr./June 1991.

MCGILCHRIST, C. A. REML Estimation for survival models with frailty. **Biometrics**, Washington, v. 49, n. 1, p. 221-225, Jan./Mar. 1993.

METROPOLIS, N.; ROSEMBLUT, A. W.; ROSEMBLUT, M. N.; TELLER, A. H.; TELLER, E. Equations of state calculations by fast computing machines. **Journal of Chemical Physics**, New York, v. 21, n. 1, p. 1087-1092, Jan./Dec. 1953.

NOGUEIRA, D. A.; SÁFADI, T.; FERREIRA, D. F. Avaliação de critérios de convergência para o método de Monte Carlo via Cadeias de Markov. **Revista Brasileira de Estatística**, Rio de Janeiro, v. 65, n. 224, p. 59-88, mar. 2004.

OAKES, D. A model for association in bivariate survival data. **Journal of the Royal Statistical Society**, London, v. 44, n. 1, p. 414-422, Jan./Mar. 1982.

PARNER, E. T. A composite likelihood approach to multivariate survival data. **Scandinavian Journal of Statistics**, Stockholm, v. 28, n. 1, p. 295-302, Jan./Mar. 2001.

PASCOA, M. A. R. de. **Intervalos de credibilidade para a razão de riscos do modelo de Cox, considerando estimativas pontuais bootstrap**. 2008. 225p. Dissertação (Mestrado em Fonoaudiologia) – Universidade Federal de Lavras, Lavras.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003. 446 p.

PEÑA, E. A. Dynamic modeling and statistical analysis of event data. **Statistical Science**, Hayward, v. 21, n. 4, p. 487-500, Jan./Mar. 2006.

PEÑA, E. A.; HOLLANDER, M. Models for Recurrent Events in Reliability and Survival Analysis. In: SOYER, R.; MAZZUCHI, T.; SINGPURWALLA, N. (Ed.). **Mathematical Reliability: an expository perspective**. Kluwer: Mathematical Reviews, 2004. p. 105–118.

PEÑA, E. A.; SLATE, E. H.; GONZALÉZ, J. R. Semiparametric inference for a general class of models for recurrent events. **Journal of Statistical Planning and Inference**, Amsterdam, v. 137, n. 6, p. 1727-1747, June 2007.

PEÑA, E. A.; SLATE, E.H.; GONZALÉZ, J. R. Semiparametric inference for a general class of models for recurrent events. **Technical Report**, South Carolina, n. 214, p.-39, Dec. 2003.

PEPE, M. S.; CAI, J. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. **Journal of the American Statistical Association**, New York, v. 88, n. 423, p. 811-820, Sept. 1993.

PETERSEN, J. H. An additive frailty model for correlated life times. **Biometrics**, Washington, v. 54, n. 3, p. 646-661, July/Sept. 1998.

PRENTICE, R. L.; WILLIAMS, B. J.; PETERSON, A. V. On the regression analysis of multivariate time data. **Biometrika**, London, v. 68, n. 2, p. 373-79, Aug. 1981.

QIOU, Z.; RAVINSHANKER, N. E.; DEY, D. K. Multivariate survival analysis with positive stable frailties. **Biometrics**, Washington, v. 55, n. 3, p. 637-644, July/Sept. 1999.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. Viena: R Foundation for Statistical Computing, 2004. Disponível em: <<http://www.R-project.org>>. Acesso em: 20 dez. 2009.

RAFTERY, A. L.; LEWIS, S. How many iterations in the Gibbs sampler? In: BERNARDO, J. M.; BERGER, J. O.; DAWID, A. P.; SMITH, A. F. M. (Ed.). **Bayesian statistics**. Oxford: University, 1992. p. 763-774.

REIS, R. L.; MUNIZ, J. A.; SILVA, F. F.; SAFADI, T.; AQUINO, L. H. Abordagem bayesiana da sensibilidade de modelos para o coeficiente de endogamia. **Ciência Rural**, Santa Maria, v. 39, n. 3, p. 1752-1759, ago./dez. 2009.

RIPATTI, S.; PALMGREN, J. Estimation of multivariate frailty models using Penalized partial likelihood. **Biometrics**, Washington, v. 56, n. 3, p. 1016-1022, July/Sept. 2000.

SAVIAN, T. V.; MUNIZ, J. A.; SAFADI, T.; SILVA, F. F. Análise bayesiana para modelos de degradabilidade ruminal. **Ciência Rural**, Santa Maria, v. 39, n. 3, p. 2169-2177, ago./dez. 2009.

SOCIEDADE BRASILEIRA DE NEFROLOGIA. São Paulo, 2009. Disponível em: <<http://www.sbn.org.br/>>. Acesso em: 29 jul. 2009.

SCHWARZ, G. Estimating the dimensional of a model. **Annals of Statistics**, Hayward, v. 6, n. 2, p. 461-464, Apr./June 1978.

SILVA, G. L. **Análise bayesiana de modelos de sobrevivência com fragilidade**. 2001. 184 p. Tese (Doutorado em Matemática) - Universidade Técnica de Lisboa, Lisboa.

SILVA, M. I. de S. **Mapeamento de QTLs em populações com pedigrees complexos**. 2008. 104 p. Tese (Doutorado em Agronomia) - Universidade Federal de Lavras, Lavras.

SINHA, D. Semiparametric bayesian analysis of multiple event time data. **Journal of the American Statistical Association**, New York, v. 88, n. 423, p. 979-983, Sept. 1993.

SINHA, D.; DEY, D. K. Semiparametric Bayesian analysis of survival data. **Journal of the American Statistical Association**, New York, v. 92, n. 439, p. 1195-1211, Dec. 1997.

STEFANESCU, C.; TURNBULL, B. W. Multivariate frailty models for exchangeable survival data with covariates. **Technometrics**, Washington, v. 48, n. 3, p. 411-417, July/Sept. 2006.

STOCKER, R. S.; PEÑA, E. A. A general class of parametric models for recurrent event data. **Technometrics**, Washington, v. 49, n. 2, p. 210-220, Apr./June 2007.

SU, X.; FAN, J. Multivariate survival trees: a maximum likelihood approach based on frailty models. **Biometrics**, Washington, v. 60, n. 1, p. 93-99, Jan./Mar. 2004.

THERNEAU, T. M.; GRAMBSCH, P. M. **Modeling survival data: extending the cox model**. New York: Springer-Verlag, 2000. 350 p.

THERNEAU, T.; HAMILTON, S. A rhDNase as an example of recurrent event analysis. **Statistics in Medicine**, New York, v. 16, n. 18, p. 2029-2047, Oct. 1997.

TOMAZELLA, V. L. D. **A modelagem de dados de eventos recorrentes via processo de Poisson com termo de fragilidade**. 2003. 165 p. Tese (Doutorado em Ciências da Computação e Matemática Computacional) - Universidade de São Paulo, São Carlos.

VAUPEL, J. W.; MATON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. **Demography**, Chicago, v. 16, p. 439-54, Jan./Mar. 1979.

WEI, L. J.; LIN, D. Y.; WEISSFELD, L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. **Journal of the American Statistical Association**, New York, v. 84, n. 408, p. 1065-1073, Dec. 1989.

WHITTMORE, G. A.; LEE, M. L. T. A multivariate survival distribution generate by inverse gaussian mixture of exponential. **Technometrics**, Washington, v. 33, n. 1, p. 39-50, Jan./Mar. 1991.

YE, Y.; KALBFLEISH, J. D.; SCHAUBEL, D. E. Semiparametric analysis of correlated recurrent and terminal events. **Biometrics**, Washington, v. 63, n. 1, p. 78-87, Jan./Mar. 2007.

## ANEXOS

ANEXO A	Página
FIGURA 1A Traço da cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_0$ , o intercepto.....	118
FIGURA 2A Traço da cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_3$ , referente à covariável raça..	119
FIGURA 3A Traço da cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_4$ , referente à covariável idade.....	119
FIGURA 4A Traço da cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_5$ , referente à covariável procedência.....	120
FIGURA 5A Traço da cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_6$ , referente à covariável nefrite.....	120
FIGURA 6A Traço da cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_7$ , referente à covariável diabetes.....	121
FIGURA 7A Traço da cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_8$ , referente à covariável hipertensão.....	121
FIGURA 8A Traço da cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_9$ , referente à covariável rim policístico.....	122
FIGURA 9A Traço da cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_{10}$ , referente à covariável outras doenças.....	122

ANEXO B	Página
FIGURA 1B Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_0$ , o intercepto.....	123
FIGURA 2B Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_1$ , referente à covariável raça.....	124
FIGURA 3B Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_2$ , referente à covariável idade.....	124
FIGURA 4B Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_3$ , referente à covariável procedência.....	125
FIGURA 5B Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_4$ , referente à covariável nefrite.....	125
FIGURA 6B Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_5$ , referente à covariável diabetes.....	126
FIGURA 7B Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_6$ , referente à covariável hipertensão.....	126
FIGURA 8B Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_7$ , referente à covariável rim policístico...	127
FIGURA 9B Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_{10}$ , referente à covariável outras doenças.	127

ANEXO C	Página
FIGURA 1C Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_0$ , o intercepto.....	128
FIGURA 2C Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_3$ , referente à covariável raça.....	128
FIGURA 3C Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_4$ , referente à covariável idade.....	129
FIGURA 4C Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_5$ , referente à covariável procedência.....	129
FIGURA 5C Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_6$ , referente à covariável nefrite.....	130
FIGURA 6C Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_7$ , referente à covariável diabetes.....	130
FIGURA 7C Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_8$ , referente à covariável hipertensão.....	131
FIGURA 8C Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_9$ , referente à covariável rim policístico..	131
FIGURA 9C Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\beta_{10}$ , referente à covariável outras doenças.....	132
FIGURA 10C Cadeia gerada pelo método MCMC e densidade a <i>posteriori</i> para o parâmetro $\delta$ da distribuição Weibull.....	132

ANEXO D	<b>Página</b>
FIGURA 1D Gráfico do logaritmo da função risco acumulado versus o tempo para a covariável sexo.....	133
FIGURA 2D Gráfico do logaritmo da função risco acumulado versus o tempo para a covariável nefrite, em que ---- representa a curva para aqueles que possuem a doença e __ representa a curva para aqueles que não a possuem.....	134
FIGURA 3D Gráfico do logaritmo da função risco acumulado versus o tempo para a covariável hipertensão, em que ---- representa a curva para aqueles que possuem a doença e __ representa a curva para aqueles que não a possuem.....	135

TABELA 8	Dados de indivíduos ( $i$ ) em tratamento de hemodiálise, contendo o número de recorrência de evento para cada indivíduo ( $D_i$ ), os tempos de ocorrência de eventos (tempos) e as covariáveis em estudo.....	136
----------	---	-----

## ANEXO A

Os gráficos do traço da cadeia e da densidade para vetor de parâmetros  $\beta$ , para a classe geral de modelos para eventos recorrentes com fragilidade multiplicativa proposta por Peña e Hollander (2004), são apresentados nas FIGURAS a seguir.

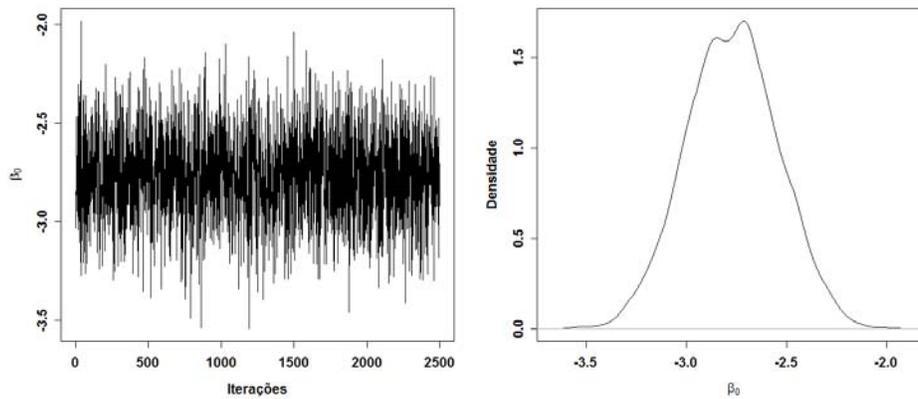


FIGURA 1A Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_0$ , o intercepto.

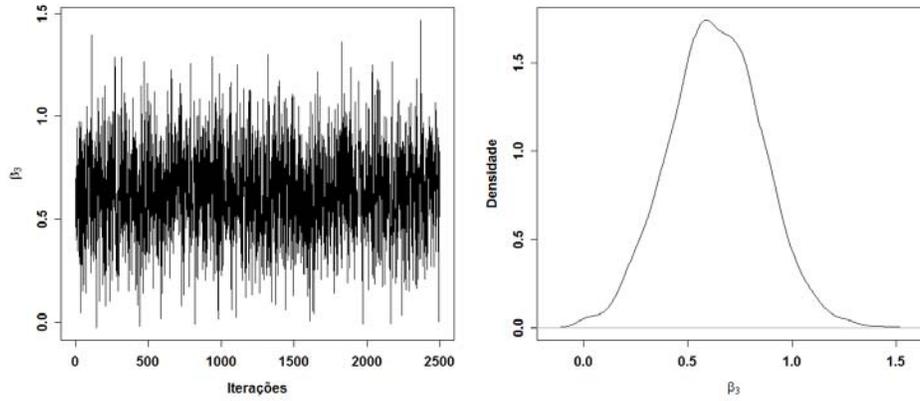


FIGURA 2A Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_3$ , referente à covariável raça.

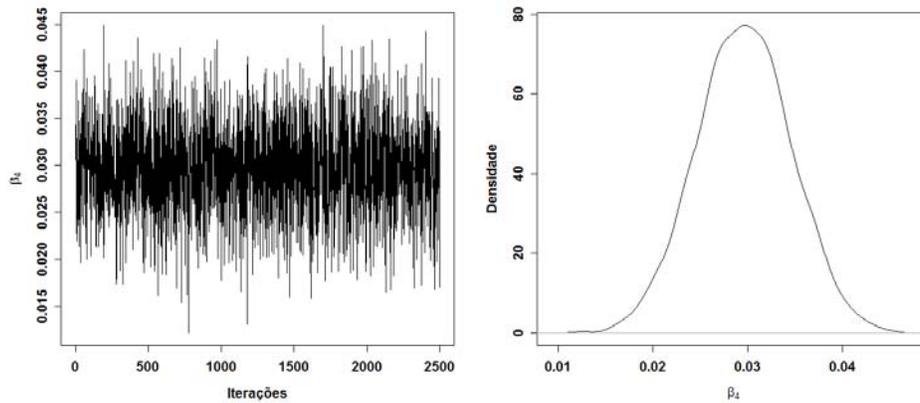


FIGURA 3A Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_4$ , referente à covariável idade.

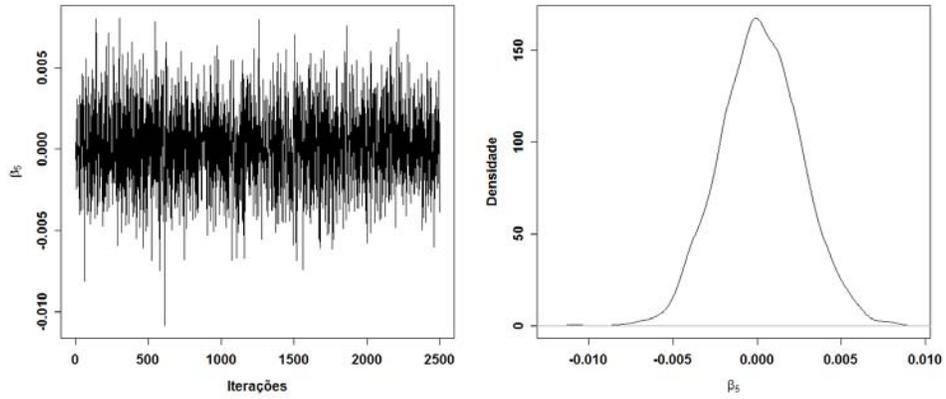


FIGURA 4A Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_5$ , referente à covariável procedência.

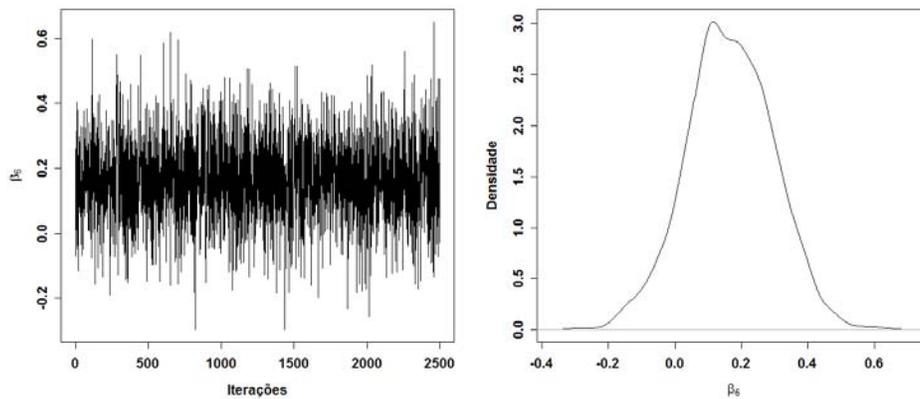


FIGURA 5A Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_6$ , referente à covariável nefrite.

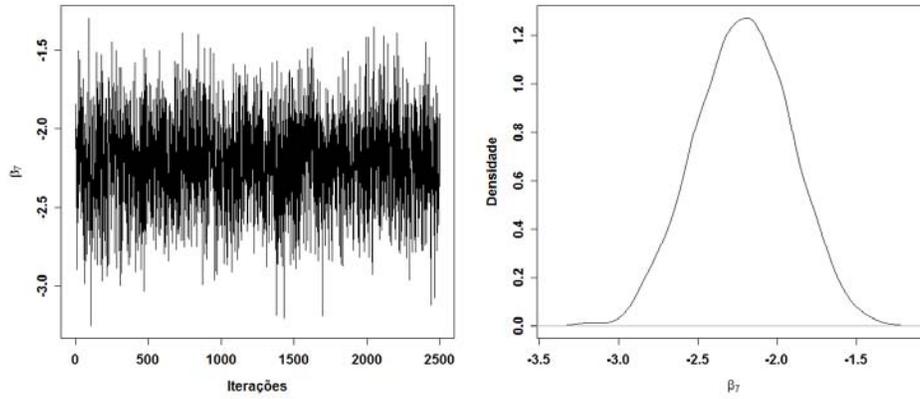


FIGURA 6A Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_7$ , referente à covariável diabetes.

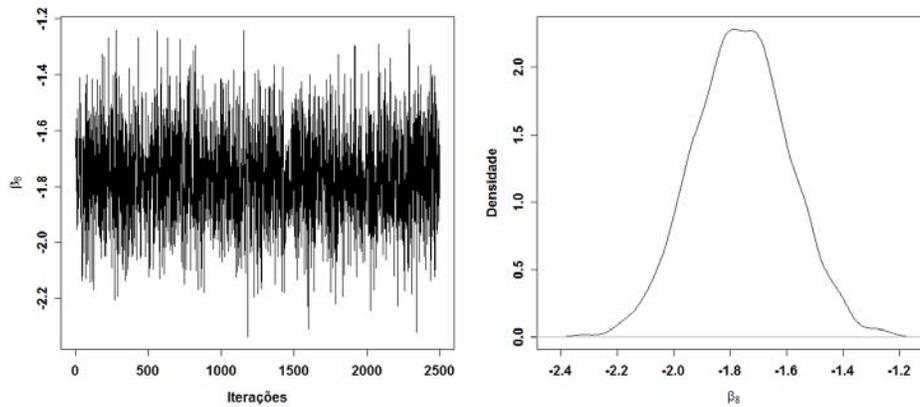


FIGURA 7A Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_8$ , referente à covariável hipertensão.

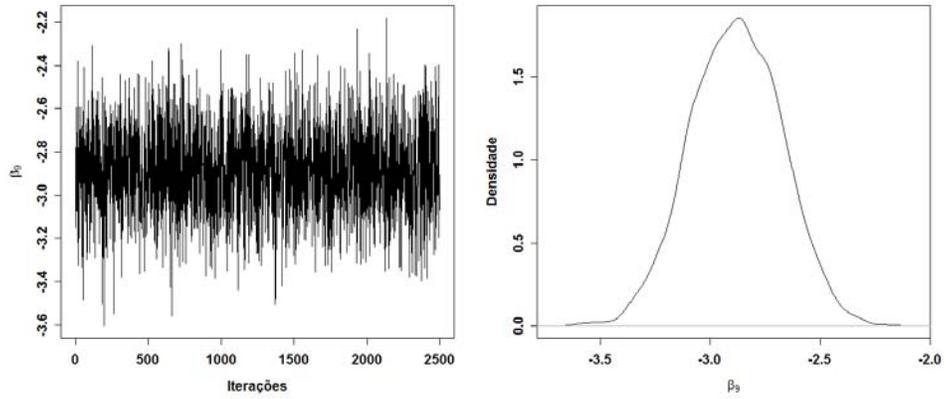


FIGURA 8A Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_9$ , referente à covariável rim policístico.

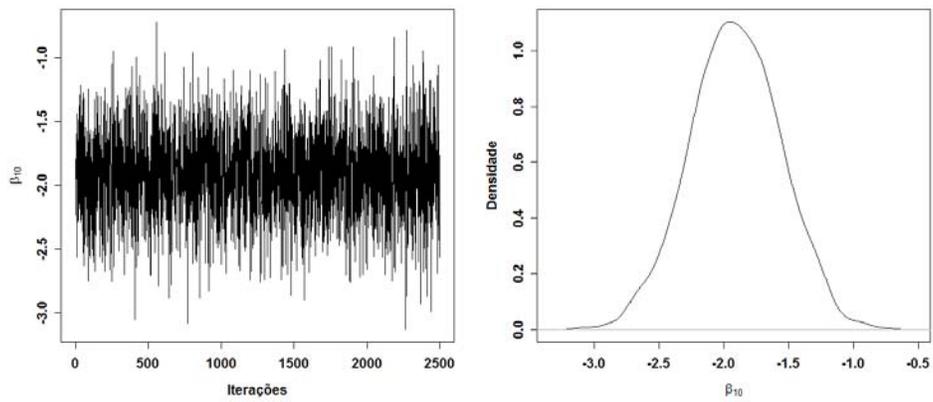


FIGURA 9A Traço da cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_{10}$ , referente à covariável outras doenças.

## ANEXO B

Os gráficos do traço da cadeia e da densidade para vetor de parâmetros  $\beta$ , para o modelo de fragilidade multiplicativo proposto por Clayton (1978), são apresentados nas FIGURAS a seguir.

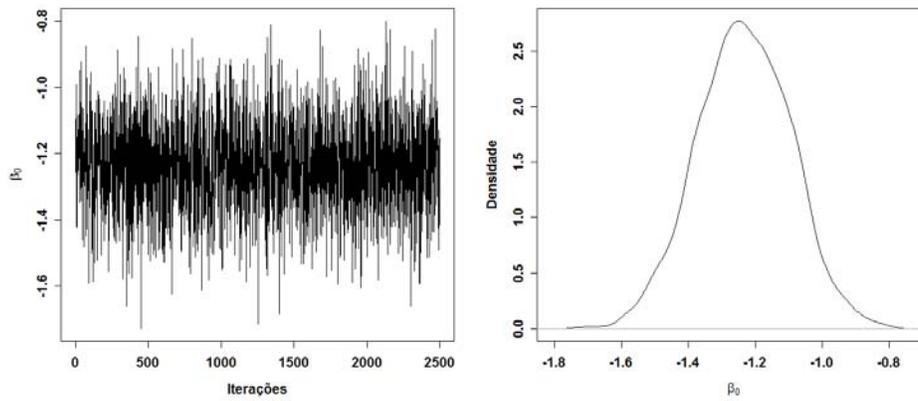


FIGURA 1B Cadeia gerada pelo método MCMC e densidade *a posteriori* para o parâmetro  $\beta_0$ , o intercepto.

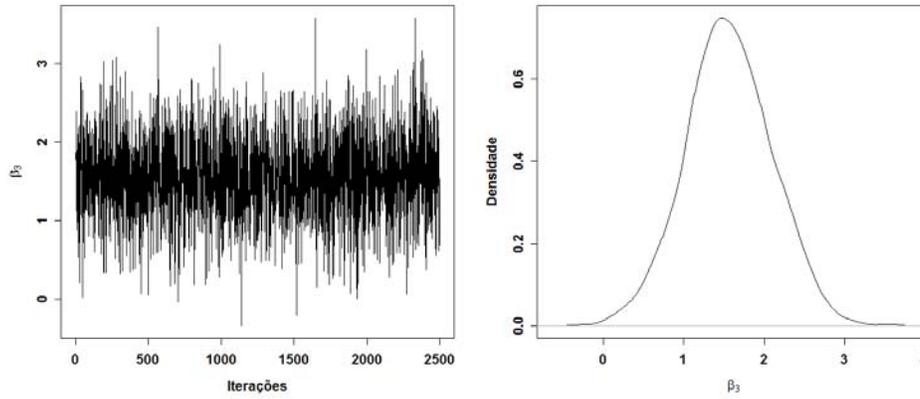


FIGURA 2B Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_3$ , referente à covariável raça.

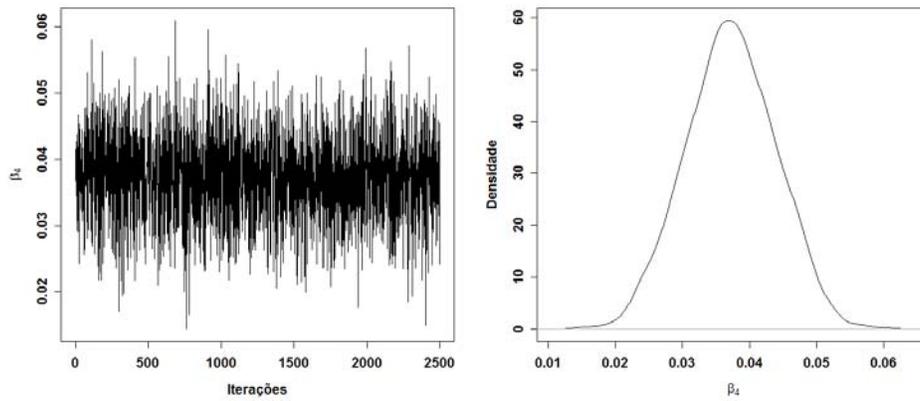


FIGURA 3B Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_4$ , referente à covariável idade.

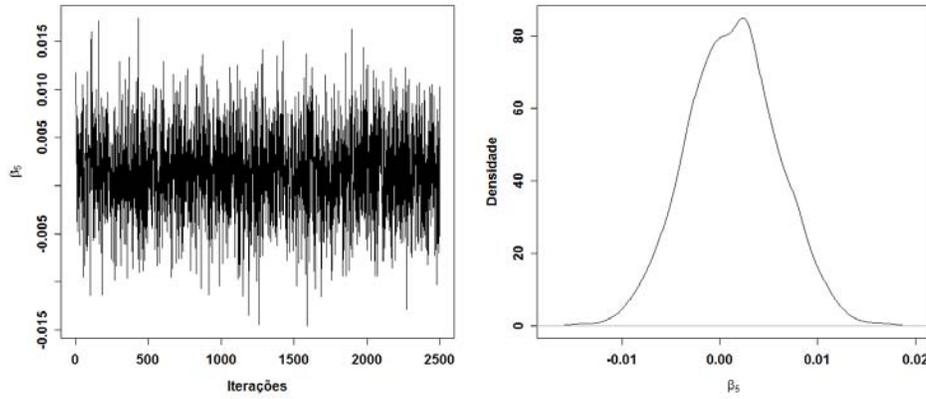


FIGURA 4B Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_5$ , referente à covariável procedência.

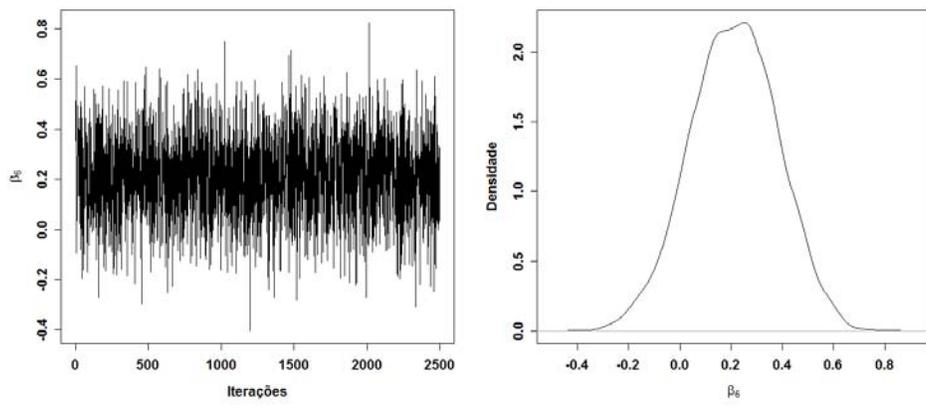


FIGURA 5B Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_6$ , referente à covariável nefrite.

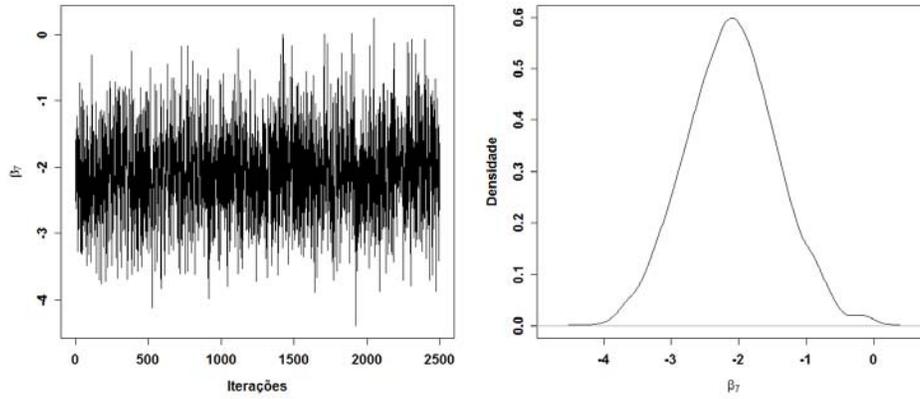


FIGURA 6B Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_7$ , referente à covariável diabetes.

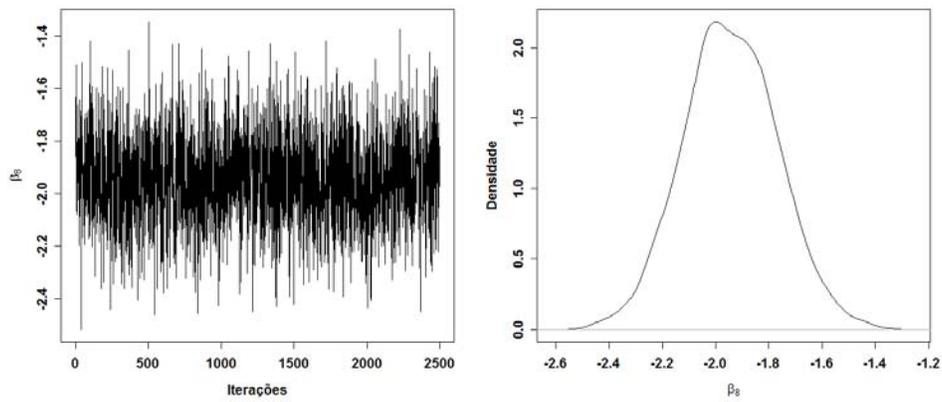


FIGURA 7B Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_8$ , referente à covariável hipertensão.

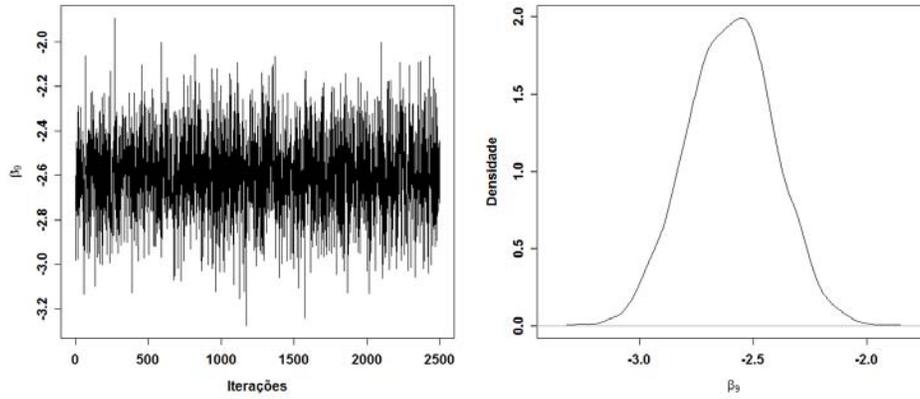


FIGURA 8B Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_9$ , referente à covariável rim policístico.

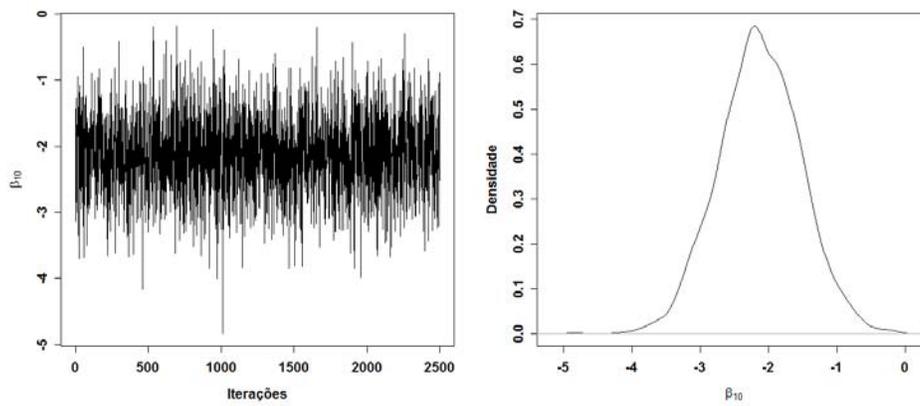


FIGURA 9B Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_{10}$ , referente à covariável outras doenças.

## ANEXO C

Os gráficos do traço da cadeia e da densidade para vetor de parâmetros  $\beta$ , para o modelo de Cox, proposto por Cox (1972) e para o parâmetro  $\beta_0$  são apresentados nas FIGURAS a seguir.

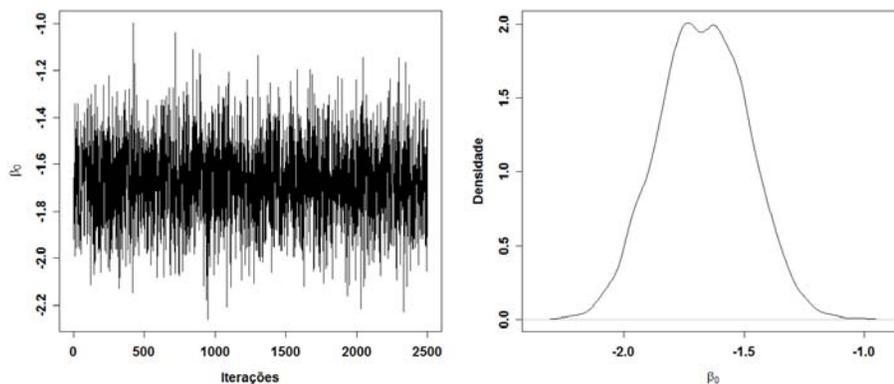


FIGURA 1C Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_0$ , o intercepto.

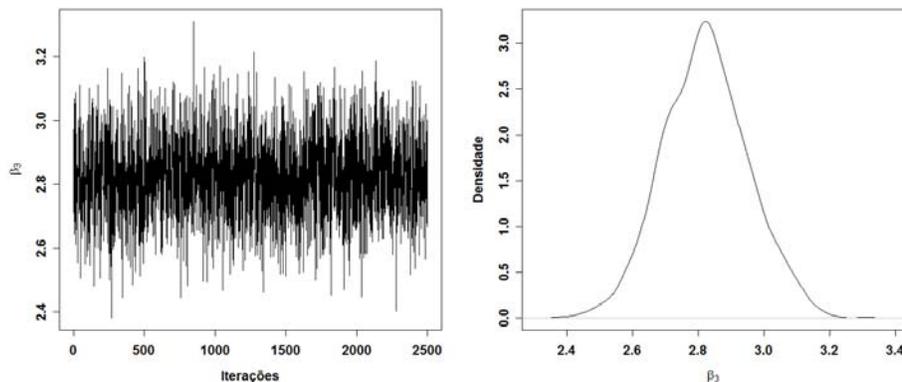


FIGURA 2C Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_3$ , referente à covariável raça.

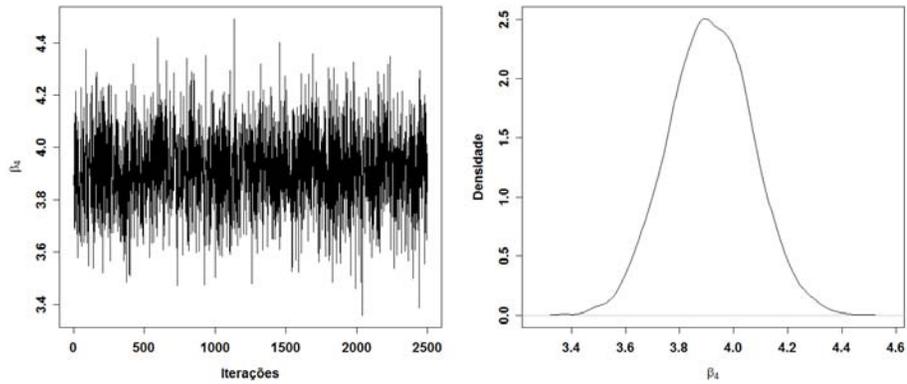


FIGURA 3C Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_4$ , referente à covariável idade.

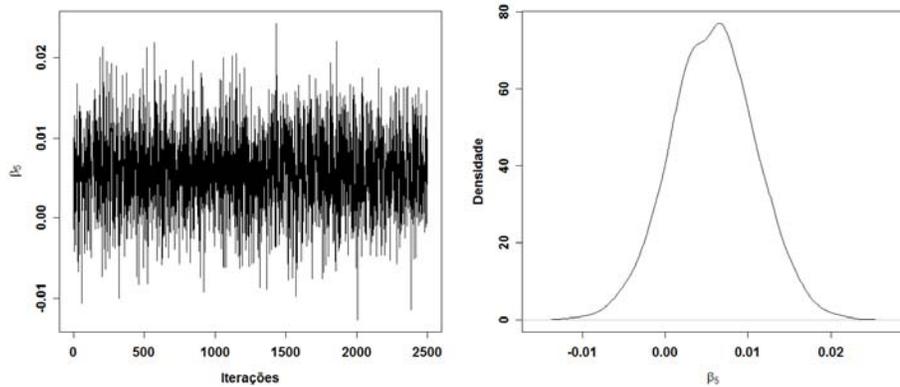


FIGURA 4C Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_5$ , referente à covariável procedência.

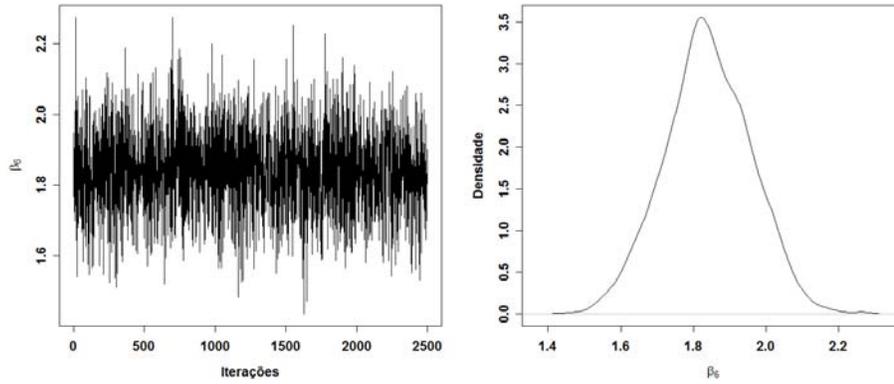


FIGURA 5C Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_6$ , referente à covariável nefrite.

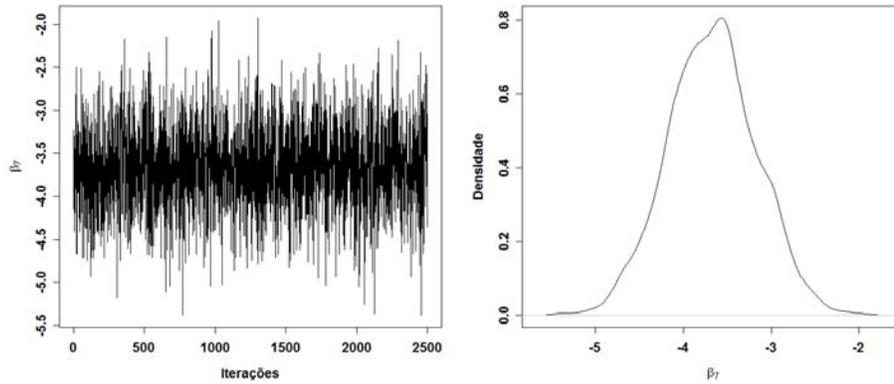


FIGURA 6C Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_7$ , referente à covariável diabetes.

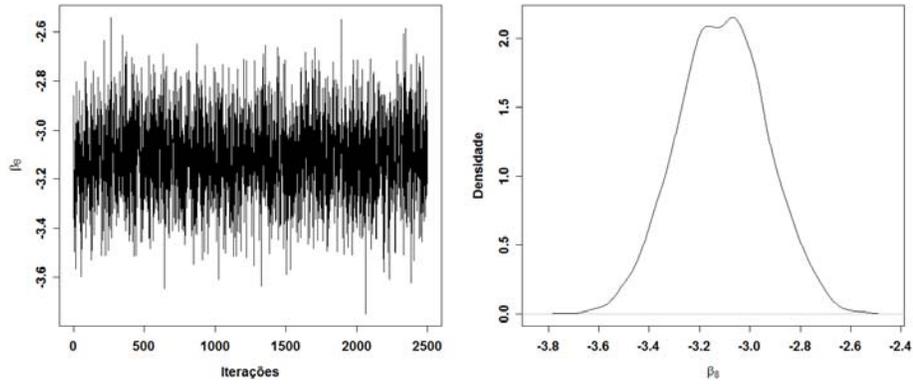


FIGURA 7C Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_3$ , referente à covariável hipertensão.

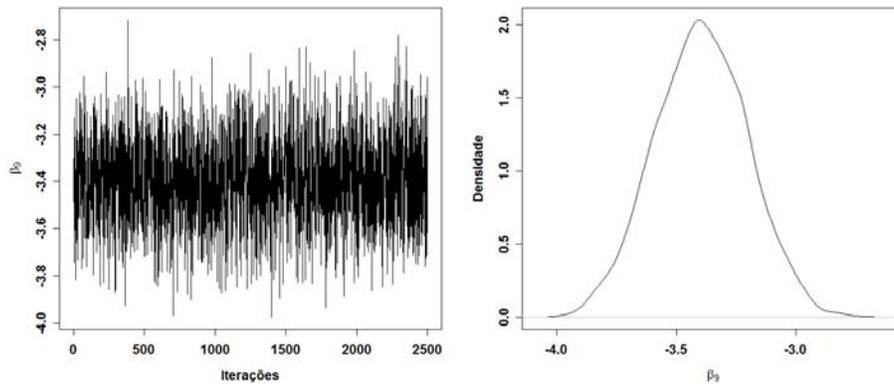


FIGURA 8C Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_5$ , referente à covariável rim policístico.

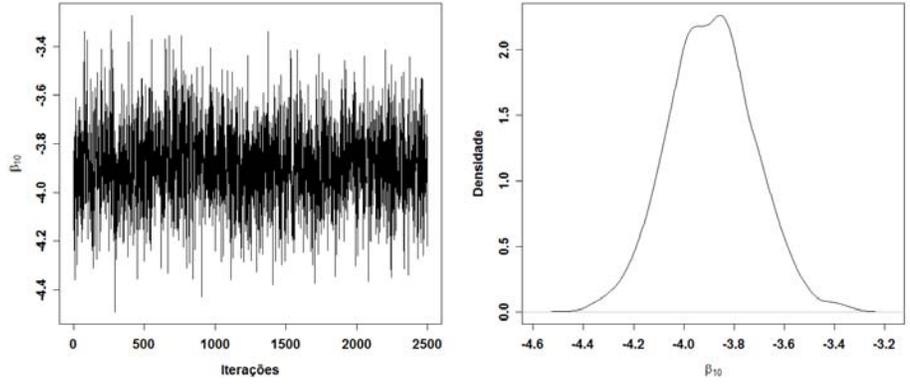


FIGURA 9C Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\beta_{10}$ , referente à covariável outras doenças.

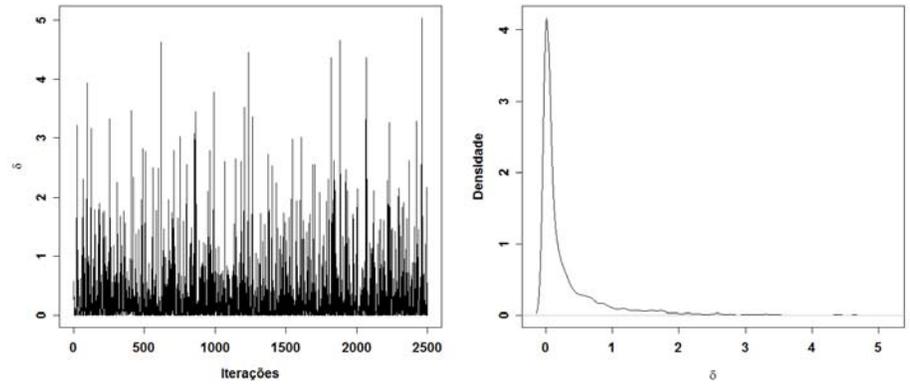


FIGURA 10C Cadeia gerada pelo método MCMC e densidade a *posteriori* para o parâmetro  $\delta$  da distribuição Weibull.

## ANEXO D

Verificação da suposição de riscos proporcionais para a utilização do modelo de regressão de Cox para as covariáveis sexo, nefrite e hipertensão, por meio dos gráficos do logaritmo da função risco acumulado versus o tempo apresentados nas FIGURAS a seguir. Os gráficos para as demais covariáveis foram suprimidos.

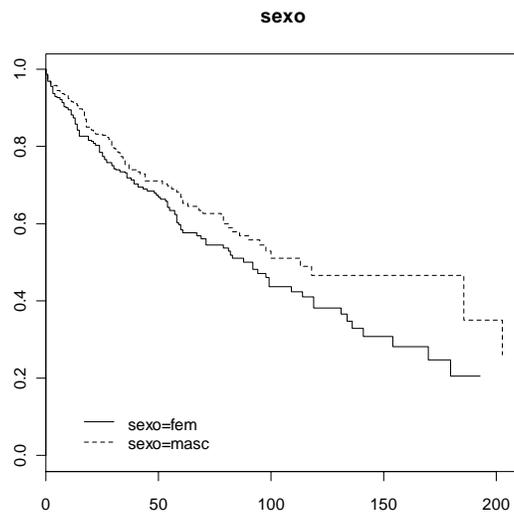


FIGURA 1D Gráfico do logaritmo da função risco acumulado versus o tempo para a covariável sexo.

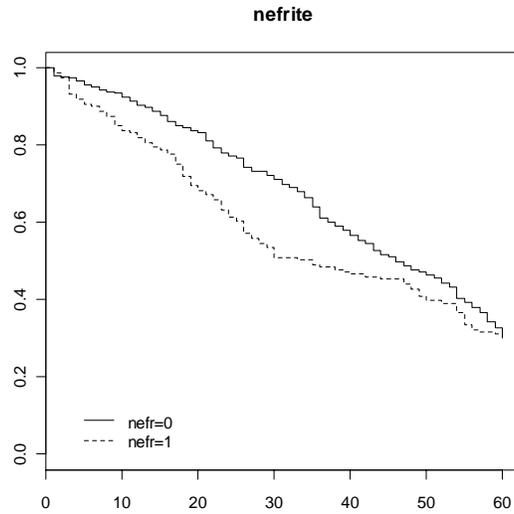


FIGURA 2D Gráfico do logaritmo da função risco acumulado versus o tempo para a covariável nephrite, em que ---- representa a curva para aqueles que possuem a doença e \_\_ representa a curva para aqueles que não a possuem.

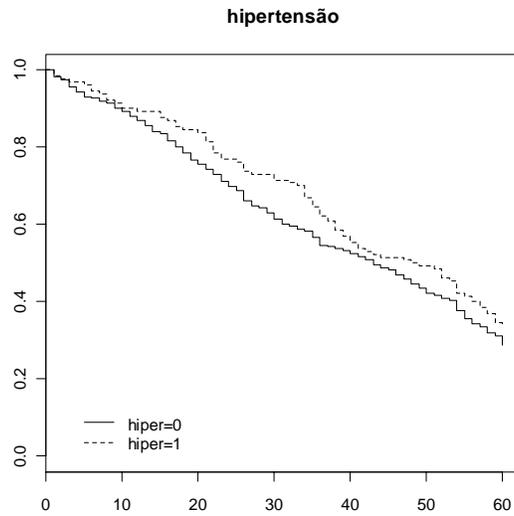


FIGURA 3D Gráfico do logaritmo da função risco acumulado versus o tempo para a covariável hipertensão, em que ---- representa a curva para aqueles que possuem a doença e \_\_ representa a curva para aqueles que não a possuem.

## ANEXO E

Na TABELA 8 é apresentada parte dos dados utilizados na aplicação desta tese, referente aos pacientes com insuficiência renal crônica (IRC) tratados no Centro de Terapia Renal Substitutiva da Santa Casa de Misericórdia de Lavras, em Lavras, MG.

TABELA 8 Dados de indivíduos ( $i$ ) em tratamento de hemodiálise, contendo o número de recorrência de evento para cada individuo ( $D_i$ ), os tempos de ocorrência de eventos (tempos) e as covariáveis em estudo.

$i$	$D_i$	tempos	Sex	Est	Raç	...	Nef	Diab	Outras
1	0	60+	0	1	0	...	0	1	0
2	0	60+	1	0	1	...	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮
32	1	58,60+	0	1	0	...	0	1	0
33	3	56,58,60+	0	0	1	...	0	0	0
34	5	50, 52, 54, 55, 56,60+	1	1	0	...	1	0	0
35	0	60+	1	0	0	...	0	0	0
36	3	43, 54,60+	1	0	0	...	0	0	0
37	1	39,60+	1	0	1	...	0	0	0
38	3	30,36,40,60+	0	0	0	...	0	1	0
39	5	43,46,47,53,58,60+	1	0	0	...	0	0	1
40	6	3,26,27,30,39,60+	0	0	1	...	1	0	0
41	4	24,25,26,40,60+	1	0	1	...	1	0	0
42	0	60+	1	0	1	...	0	0	0
43	1	36,60+	0	0	0	...	1	0	0
44	0	60+	1	1	1	...	0	0	1

⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
113	5	3,5,22,23,30,60+	0	0	1	⋮	1	0	0
114	5	9,17,18,23,60	1	0	0	⋮	1	0	0
115	6	5,11,17,26,27,56,60+,	0	1	0	⋮	0	0	0
116	6	17,18,21,22,26,30,60+	0	1	0	⋮	0	0	0
117	1	56,60+	1	0	0	⋮	1	0	0
118	1	39,60+	0	1	0	⋮	0	1	0
119	3	12,13,22,60+	1	0	1	⋮	0	1	0
120	11	3,6,11,12,13,14,15,16,18, 30,60+	1	1	0	⋮	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
149	0	60+	0	0	0	⋮	0	1	0

O símbolo “+” indica que, no tempo final de estudo, ou seja, na última observação do paciente, não foi verificada nova ocorrência de evento.

As cováriaveis foram dicotomizadas da seguinte maneira:

1. sexo: 0 se feminino ou 1, se masculino;
2. estado civil do paciente: 0 se casado ou 1, se solteiro/viúvo/separado;
3. raça: 0 se raça branca/parda ou 1, se raça negra;
4. idade no início do tratamento (em anos);
5. procedência: 0 se microrregião de Lavras ou 1, se outras regiões;
6. nefrite: 0 se o paciente não possui a doença ou 1, se possui;
7. diabetes *mellitus*: 0 se o paciente não possui a doença ou 1, se possui;
8. hipertensão: 0 se o paciente não possui a doença ou 1, se possui;
9. presença de rim policístico: 0 se o paciente não possui ou 1, se possui;
10. outras doenças: 0 se o paciente não possui a doença ou 1, se possui.