UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

Modelagem de Dados de Sobrevivência com Eventos Recorrentes via Fragilidade Discreta

Márcia Ap. Centanin Macera

São Carlos - SP
Outubro/2015

Márcia Ap. Centanin Macera

Modelagem de Dados de Sobrevivência com Eventos Recorrentes via Fragilidade Discreta

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

Orientador: Prof. Dr. Francisco Louzada Neto
Co-orientador: Prof. Dr. Vicente Garibay Cancho

São Carlos - SP
Outubro/2015

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária UFSCar
Processamento Técnico
com os dados fornecidos pelo(a) autor(a)

M142m

Macera, Márcia Aparecida Centanin
Modelagem de dados de sobrevivência com eventos recorrentes via fragilidade discreta / Márcia Aparecida Centanin Macera. -- São Carlos : UFSCar, 2015.
110 p.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2015.

1. Análise de sobrevivência. 2. Eventos recorrentes. 3. Processo de Poisson. 4. Fragilidade discreta. 5. Máxima verossimilhança. I. Título.

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado da candidata Marcia Aparecida Centanin Macera, realizada em 02/09/2015:

Prof. Dr. Francisco Louzada Neto
USP

Prof. Dr. Adriano Kamimura Suzuki
USP

Prof. Dr. Jorge Luis Bazán Guzmán
USP

Profa. Dra. Teresa Cristina Martins Dias
UFSCar

Profa. Dra. Vera Lucia Damasceno Tomazella
UFSCar

Prof. Dr. Vicente Garibay Cancho
USP

Agradecimentos

A Deus pela saúde e força para superar as dificuldades.

Aos meus orientadores, Prof. Dr. Francisco Louzada Neto e Prof. Dr. Vicente Garibay Cancho, por todo empenho, sabedoria e compreensão com que me conduziram durante todo esse tempo. Agradeço ainda pelo conhecimento que adquiri com vocês e por todo incentivo no decorrer do trabalho.

Aos Professores, Dr. Mário de Castro e Dr. Enrico Colosimo, pelas sugestões e correções importantes à melhoria do trabalho.

Ao Prof. Dr. Cor Jesus Fernandes Fontes que forneceu gentilmente um conjunto de dados para análise.

Agradeço especialmente aos meus pais, que sempre me estimularam a prosseguir os estudos e me deram conforto para cumpri-los. Exemplos de vida e honestidade.

Ao meu noivo Rafael, pelo incentivo, amor, compreensão e atenção. Agradeço por ter sempre uma palavra de conforto e estímulo nos momentos difíceis. Obrigada também por toda ajuda nas correções e, acima de tudo, por sempre estar ao meu lado.

Aos funcionários do Departamento de Estatística da UFSCar, especialmente à Maria Isabel de Araujo, pela dedicação e convívio.

À todos aqueles que, direta ou indiretamente, tive o prazer de trocar conhecimentos para que essa tese atingisse seus objetivos.

Finalmente, agradeço à CAPES pelo apoio financeiro durante o desenvolvimento desse trabalho.

Resumo

Neste trabalho propomos metodologias alternativas e extensões em modelos para dados de eventos recorrentes. Especificamente, propomos um modelo em que a distribuição condicional do tempo entre sucessivas ocorrências de um evento recorrente é derivada facilmente da função de taxa marginal, proporcionando interpretações práticas mais diretas, além de considerar a relação entre as sucessivas ocorrências para cada indivíduo. O outro modelo, que estende os modelos de fragilidade para dados de eventos recorrentes permitindo o uso de distribuições como Bernoulli, Geométrica, Poisson, Weibull Discreta, Binomial Negativa ou outra distribuição discreta para a variável de fragilidade, também foi proposto. O procedimento de estimação dos parâmetros para ambos modelos foi realizado considerando-se o método de máxima verossimilhança. Estudos de simulação foram realizados com o objetivo de analisar algumas propriedades frequentistas do método de estimação e avaliar a qualidade das estimativas de máxima verossimilhança. Aplicações a conjuntos de dados reais mostraram a aplicabilidade dos modelos propostos. De modo geral, os modelos propostos mostraram-se adequados para a modelagem de dados de eventos recorrentes.

Palavras-chave: Análise de sobrevivência; Eventos recorrentes; Processo de Poisson; Fragilidade discreta; Máxima verossimilhança

Abstract

In this thesis it is proposed alternative methodologies and extensions on models for recurrent event data. Specifically, we propose a model in which the distribution of the gap time is easily derived from the marginal rate function providing more direct practical interpretation besides to consider the relation between successive gap times for each individual. Another model that extends the frailty models for recurrent event data to allow a Bernoulli, Geometric, Poisson, Discrete Weibull, Negative Binomial or other discrete distribution of the frailty variable has also been proposed. The parameter estimation procedure for both models was conducted considering maximum likelihood methods. Simulation studies were performed in order to examine some frequentist properties of the estimation method and evaluate the maximum likelihood estimates quality. Real data applications demonstrated the use of the proposed models. Overall, the proposed models were suitable for analyzing recurrent event data.

Keywords: Survival analysis; Recurrent events; Poisson process; Discrete Frailty; Maximum likelihood

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A análise de sobrevivência é formada pelo conjunto de técnicas estatísticas utilizadas para a análise de dados que envolvem tempos, podendo ser tempos de vida de indivíduos, itens ou componentes. Os estudos em análise de sobrevivência e confiablidade envolvem o acompanhamento de indivíduos (unidades), cujo foco é o tempo até a ocorrência de um determinado evento de interesse, denominado tempo de falha ou tempo de ocorrência (Collett, 2003). O evento de interesse pode ser a morte de um indivíduo, o aparecimento de um tumor, a falha de um componente ou equipamento eletrônico, entre outros. Com exceção do primeiro exemplo, os demais podem ocorrer diversas vezes para um mesmo indivíduo, denominados eventos recorrentes. Dados de eventos recorrentes são predominantes em uma ampla variedade de situações, como em estudos médicos, confiabilidade e engenharia, criminologia e demografia. A análise estatística dos dados de eventos recorrentes é distinta das análises usuais para dados de sobrevivência, e ignorar a recorrência do evento pode comprometer a eficácia da metodologia. Nesse sentido, diferentes metodologias têm sido propostas com o objetivo de analisar os dados de eventos recorrentes (Cook & Lawless, 2007).

Quando os indivíduos estão sujeitos à múltiplas ocorrências do mesmo evento, eventos recorrentes, a suposição de independência entre os tempos

de sobrevivência pode não ser válida. Como se observa mais de um tempo para cada indivíduo, é razoável supor que exista uma dependência entre os mesmos. Um modelo que geralmente é muito utilizado na modelagem dessa dependência entre os tempos de recorrência de um indivíduo é o modelo de fragilidade (Hougaard, 2000). Os modelos de fragilidade são caracterizados pela introdução de um efeito aleatório, representado por uma variável aleatória contínua não observável, no modelo. Este efeito aleatório, denominado fragilidade, representa as informações que não podem ou não foram observadas e tem como objetivo explicar a correlação entre os tempos multivariados. A distribuição paramétrica mais assumida para a fragilidade é a distribuição Gama, que se deve ao fato de tal distribuição ser não negativa, flexível e algebricamente conveniente. No entanto, outras distribuições também têm sido utilizadas. Por outro lado, uma vez que a fragilidade é não observável, a escolha da distribuição de fragilidade adequada pode ser um problema. Com base nisso, em uma primeira etapa, propomos um modelo de taxa marginal para analisar tempos entre eventos recorrentes, que é formulado considerando cada um dos tempos entre eventos condicionado ao tempo da recorrência anterior. Nessa formulação, os tempos entre eventos são tratados da mesma forma e a relação entre as sucessivas ocorrências deixa de ser um problema. Além disso, possibilita interpretações práticas mais diretas para a identificação de fatores de riscos.

Motivados pelos avanços dos tratamentos médicos, pesquisadores passaram a estudar a possibilidade de um indivíduo deixar de ser suscetível a um evento de interesse. Nesse sentido, as distribuições de fragilidade contínuas não permitem a possibilidade de um indivíduo apresentar risco zero para a ocorrência de um evento. Assim, em uma segunda etapa, propomos um modelo para dados de eventos recorrentes que estende os modelos com fragilidade permitindo o uso de distribuições como Bernoulli, Geométrica, Poisson ou outras distribuições discretas para a variável de fragilidade. A fragilidade neste

caso pode representar, por exemplo, a presença de um número desconhecido de fatores que levam à recorrência do evento. Ainda, fragilidade zero corresponde a um modelo que contém uma proporção de indivíduos que nunca falham.

Para todos os modelos propostos foram realizados estudos de simulação com o objetivo de analisar algumas propriedades frequentistas do procedimento de estimação. Aplicações a um conjunto de dados reais da literatura referente a sucessivas reinternações de pacientes diagnosticados com câncer colorretal mostraram a aplicabilidade dos modelos propostos e permitiram uma comparação entre os resultados obtidos nesta tese e aqueles obtidos anteriormente na literatura a partir de modelagens similares. As metodologias propostas também foram aplicadas a um conjunto de dados referente à malária, diagnosticada em pacientes atendidos pela Faculdade de Medicina da UFMT, no intuito de determinar quais covariáveis estão relacionadas com um aumento e/ou diminuição do tempo até a recorrência do evento de interesse e também estimar a probabilidade de indivíduos não suscetíveis por meio da introdução da fragilidade discreta no modelo.

1.1 Descrição dos objetivos

No contexto apresentado acima, o principal objetivo desta tese é propor metodologias alternativas e extensões em modelos para dados de eventos recorrentes, em particular estender os modelos com fragilidade permitindo o uso de distribuições como Bernoulli, Geométrica, Poisson, Weibull Discreta, Binomial Negativa ou outra distribuição discreta para a variável de fragilidade.

Dessa forma, podemos relacionar os seguintes objetivos específicos que pretende-se alcançar no decorrer desta tese:

(i) propor um modelo alternativo para os tempos entre recorrências considerando a possível correlação existente entre os tempos de um mesmo indivíduo. Análise inferencial, testes de hipóteses e resíduos para esse

modelo;

(*ii*) propor uma extensão para os modelos com fragilidade existentes para eventos recorrentes considerando uma fragilidade discreta, a partir do modelo citado no item (*i*);

(*iii*) avaliação do desempenho do procedimento de estimação dos parâmetros do modelos por meio de estudos de simulação;

(*iv*) validação das metodologias propostas por meio da aplicação dos modelos a conjuntos de dados reais.

1.2 Apresentação dos capítulos

Este primeiro capítulo apresenta a contextualização e motivação desta tese, bem como seus principais objetivos. No Capítulo 2 é apresentado um referencial teórico contendo tópicos importantes em análise de sobrevivência que embasaram a pesquisa apresentada nesta tese. Em particular, é apresentada uma revisão bibliográfica dos modelos para dados de eventos recorrentes e suas principais características, sendo este o tema abordado nesta tese. No Capítulo 3 é apresentado um modelo alternativo para analisar tempos entre eventos recorrentes. O processo de recorrência associado segue um processo de Poisson não homogêneo e a respectiva função de taxa é modelada por um estrutura multiplicativa. Uma abordagem clássica é considerada para estimação dos parâmetros do modelo, incluindo testes de hipóteses e uma ferramenta de diagnóstico para testar a adequabilidade e avaliar a qualidade do ajuste do modelo. Avalia-se a metodologia proposta através de conjuntos de dados simulados e apresenta-se uma aplicação do modelo em um conjunto de dados reais, bem como os seus resultados. Resultaram do Capítulo 3 dois artigos, Macera *et al.* (2014) e Louzada *et al.* (2015), publicados, respectivamente, nos periódicos *Biometrical Journal* e *Journal of Applied Statistics*. No Capítulo 4

é apresentado um modelo, considerando o modelo abordado no Capítulo 3, o qual é induzido por uma fragilidade discreta. Novamente, uma abordagem clássica utilizando métodos de máxima verossimilhança é considerada para fazer inferência sobre os parâmetros do modelo. Um procedimento de reamostragem *bootstrap* é considerado como uma alternativa para a obtenção de intervalos de confiança. Além disso, é apresentado um estudo de simulação com o objetivo de avaliar o desempenho do método de estimação proposto, e uma aplicação do modelo para um conjunto de dados reais da literatura. Um estudo envolvendo recorrências de malária em indivíduos atendidos pela Faculdade de Medicina da Universidade Federal de Mato Grosso (UFMT) é considerado no Capítulo 5. Os modelos abordados nos Capítulos 3 e 4 são ajustados aos dados ilustrando a abordagem proposta e a interpretação dos parâmetros. Por fim, as principais conclusões desta tese com base nos resultados obtidos, bem como algumas perspectivas futuras de pesquisa são discutidas no Capítulo 6.

Capítulo 2

Análise de Sobrevivência

Análise de sobrevivência é um termo utilizado para descrever a análise de dados que envolvem tempos, os quais podem ser tempos de vida de indivíduos, itens ou componentes (Collett, 2003). O tempo de vida, também denominado tempo de falha ou tempo de sobrevivência, é uma medida contada a partir de um tempo inicial, bem definido, até a ocorrência de um evento de interesse. Este tempo pode ser o tempo até a morte do paciente bem como até a cura ou recidiva de uma doença. De forma geral, a análise de sobrevivência pode ser definida como a análise do tempo até a ocorrência de um determinado evento. Em muitas situações, este evento não chega a ocorrer para alguns indivíduos (unidades) durante o período de observação, não sendo possível observar o tempo de sobrevivência para todos os indivíduos em estudo. Sendo assim, obtém-se apenas uma informação incompleta para esses indivíduos, dando origem às chamadas censuras. Mesmo sendo incompletas, essas observações fornecem informações sobre o tempo de sobrevivência dos indivíduos e, se forem ignoradas, podem ocasionar em inferências incorretas (Colosimo & Giolo, 2006). Nesses casos, há necessidade da inclusão de uma variável extra na análise, que indica se o valor do tempo de sobrevivência foi ou não completamente observado. De acordo com Kalbfleisch & Prentice (2002), a observação parcial da resposta e a presença de censura são as principais

características de dados de sobrevivência.

Por outro lado, alguns eventos de interesse não são terminais e podem ocorrer mais de uma vez para o mesmo indivíduo, originando assim os eventos recorrentes. Dados de eventos recorrentes surgem frequentemente em estudos longitudinais envolvendo múltiplos sujeitos e, podem ser observados em diversas áreas, tais como biomedicina, saúde pública, engenharia e confiabilidade, demografia, ciência, política e economia, entre outras. A análise estatística destes dados torna-se distinta das análises usuais, sendo de grande importância que metodologias apropriadas sejam desenvolvidas para este tipo de dados.

Apresentamos a seguir alguns tópicos importantes em análise de sobrevivência que embasaram a pesquisa apresentada nesta tese.

2.1 Conceitos básicos em análise de sobrevivência

O tempo de sobrevivência observado, t, para um indivíduo é uma realização de uma variável aleatória T, contínua e não-negativa. O comportamento desta variável pode ser expresso através de funções matemáticas equivalentes, de modo que se uma delas é especificada as demais podem ser obtidas.

Associado à variável aleatória T tem-se uma distribuição de probabilidade caracterizada por uma função densidade de probabilidade $f(t)$. Em geral, existem duas funções que são de interesse central em análise de sobrevivência, denominadas função de sobrevivência e função de risco. A função de sobrevivência é definida como a probabilidade de um indivíduo em observação não falhar (ou do evento de interesse não ocorrer), pelo menos até um instante de tempo t, e é dada por

$$S(t) = \Pr(T \geq t) = 1 - F(t), \tag{2.1}$$

em que $F(t)$ é a função de distribuição acumulada da variável aleatória T. A função de sobrevivência em (2.1) é uma função monótona não-crescente no intervalo de tempo $[0, \infty)$, tal que $S(0) = 1$ e $S(\infty) = \lim_{t\to\infty} S(t) = 0$ (Lawless, 2003).

A função de risco fornece a taxa instantânea de falha por unidade de tempo, isto é, o limite da probabilidade de um indivíduo falhar no intervalo de tempo $[t, t + \Delta t)$, com $\Delta t \to 0$, dado que o indivíduo sobreviveu até o instante t. A função de risco é expressa por

$$h(t) = \lim_{\Delta t\to 0} \frac{\Pr\{t \leq T \leq t + \Delta t \mid T \geq t\}}{\Delta t}. \tag{2.2}$$

Uma outra função de interesse quando se considera dados de sobrevivência é a função de risco acumulado, dada por

$$H(t) = \int_0^t h(u)du. \tag{2.3}$$

Existe uma relação matemática entre as funções básicas de sobrevivência $f(\cdot), S(\cdot), h(\cdot)$ e $H(\cdot)$ discutidas anteriormente, sendo que o conhecimento de uma delas nos permite o cálculo de todas as outras. A Tabela 2.1 apresenta as relações entre essas funções de interesse em análise de sobrevivência.

Tabela 2.1: Relação entre as funções básicas de sobrevivência.

Função	Função obtida			
especificada	$f(t)$	$S(t)$	$h(t)$	$H(t)$
$f(t)$	-	$\int_t^\infty f(u)du$	$\frac{f(t)}{\int_t^\infty f(u)du}$	$-\log\left(\int_t^\infty f(u)du\right)$
$S(t)$	$-\frac{dS(t)}{dt}$	-	$-\frac{d\log(S(t))}{dt}$	$-\log(S(t))$
$h(t)$	$h(t)\exp\left\{-\int_0^t h(u)du\right\}$	$\exp\left\{-\int_0^t h(u)du\right\}$	-	$\int_0^t h(u)du$
$H(t)$	$\frac{dH(t)}{dt}\exp\{-H(t)\}$	$\exp\{-H(t)\}$	$\frac{dH(t)}{dt}$	-

2.1.1 Censura

Uma das principais características de dados de sobrevivência é a presença de censura. A censura está relacionada ao fato da variável aleatória de interesse (tempo de sobrevivência) não ser completamente observada para alguns indivíduos, obtendo-se assim uma informação parcial. Isto ocorre quando o acompanhamento do indivíduo é interrompido por algum motivo, por exemplo, quando o mesmo abandona o estudo antes da observação do evento de interesse, ou o estudo termina, ou o indivíduo morre devido a uma causa diferente da estudada.

As censuras podem ocorrer de várias formas, de acordo com diferentes tipos e mecanismos. Existem basicamente três tipos de censura, sendo eles: 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, em geral, representada através de uma variável indicadora que assume valor 1 se o tempo de sobrevivência é completamente observado e 0 caso contrário (Carvalho *et al.*, 2005).

A censura do tipo I ocorre quando o estudo termina após um período de tempo pré-estabelecido, enquanto a censura do tipo II ocorre quando o estudo termina após o evento de interesse ter ocorrido um número pré-estabelecido de vezes. Já a censura aleatória ocorre quando o indivíduo abandona o estudo sem que tenha apresentado o evento de interesse.

A censura à direita é aquela em que o tempo de ocorrência do evento de interesse é maior do que o tempo registrado para o final do estudo, enquanto a censura à esquerda é aquela em que o tempo registrado de ocorrência do evento é anterior ao início do estudo. Por fim, a censura intervalar é um tipo de censura mais geral que é caracterizada pelo fato de o tempo de ocorrência do evento não ser conhecido exatamente, mas sim pertencer a um intervalo de tempo (Lawless, 2003).

É importante destacar as diferenças entre censura e truncamento. O

conceito de censura é diferente do conceito de truncamento e ambos não devem confundidos. Em amostras em que o fenômeno de censura ocorre, é realizado o registro de todos os casos, mesmo daqueles que são tidos como censurados. Por outro lado, uma amostra truncada pode ser vista como uma amostra em que todos os valores fora dos limites de truncamento são omitidos e nem mesmo um registro dos casos omitidos é mantido. A censura pode ainda ser vista como uma característica resultante do processo de coleta dos dados, enquanto o truncamento é uma característica da população da qual a amostra é coletada (Cook & Lawless, 2007).

Dessa forma, tanto a censura quanto o truncamento resultam em falta de informação sobre a variável de interesse. No entanto, é necessário destacar que a principal diferença entre estas duas características dos dados é que no caso de censura há o registro do caso não observado e no truncamento não há registro de casos não observados. Esta diferença estrutural determina se os dados serão abordados como censurados ou truncados.

2.1.2 Modelagem de dados de sobrevivência

Os dados de sobrevivência, além de incorporar os tempos de sobrevivência e a variável indicadora de censura, podem estar associados a um conjunto de outras variáveis observáveis, denominadas variáveis explicativas ou covariáveis, as quais contém informações complementares de cada indivíduo. Estas variáveis podem representar tanto características do próprio indivíduo, tais como raça, idade, sexo; como características externas ao indivíduo, tais como possíveis tratamentos aos quais o mesmo foi submetido, fatores ambientais, entre outros. Nesses casos, o interesse está geralmente centrado em estudar a forma como uma ou mais destas variáveis podem afetar o tempo de sobrevivência de um indivíduo. Quando essas situações surgem, uma maneira de investigar a influência das covariáveis nos tempos de sobrevivência é através de modelos de regressão, em que a relação do tempo de sobrevivência com a

covariável é explicitamente identificada.

Quando os tempos de sobrevivência estão relacionados com covariáveis, diz-se que a população é heterogênea. Caso contrário, a população é dita homogênea. Usualmente, a investigação da influência de covariáveis no tempo de sobrevivência, em populações heterogêneas, é feita por meio da modelagem da função de risco $h(t)$. Existe uma extensa literatura sobre as técnicas para a análise dos modelos de regressão. Collett (2003); Carvalho *et al.* (2005) discutem modelos semiparamétricos, em especial o modelo de riscos proporcionais de Cox (Cox, 1972a). Uma ênfase em modelos paramétricos pode ser encontrada em Lawless (2003). Ainda, os modelos de regressão podem ser analisados através da teoria de processos de contagem, como apresentado em Gill (1984).

O modelo introduzido por Cox (1972a) é o modelo semiparamétrico mais popular utilizado para a análise de dados de sobrevivência. Nesse modelo, é assumido que os tempos de sobrevivência são independentes, ou seja, os indivíduos não são correlacionados. Além disso, esse modelo comporta bem dados com censura além de incorporar naturalmente covariáveis dependentes do tempo. Uma breve discussão sobre este modelo é incluída para complemento do texto.

Considere um vetor de p covariáveis, $\mathbf{x} = (x_1, \ldots, x_p)$, coletadas em um tempo inicial, e seja $h_0(t)$ uma função arbitrária não-negativa do tempo, com $\mathbf{x} = (0, \ldots, 0)$. Segundo Cox (1972a), pode-se considerar a função de risco como o produto dessa função $h_0(t)$ e uma função não negativa das covariáveis. Sendo assim, o modelo de regressão multiplicativo estabelece que a função de risco de um indivíduo, no tempo t, pode ser escrita como

$$h(t|\mathbf{x}) = h_0(t)g(\mathbf{x}, \boldsymbol{\beta}), \tag{2.4}$$

em que $g(\cdot)$ é uma função não-negativa conhecida que é igual a 1 quando seu argumento é zero, e $\boldsymbol{\beta}$ é um vetor de parâmetros desconhecidos associado às

covariáveis. A função $h_0(\cdot)$ descreve a dependência do risco sobre o tempo e é denominada função de risco basal, sendo comum a todos os indivíduos. Uma variedade de formas funcionais pode ser empregada para $g(\cdot)$, em que a mais simples e natural consiste em fazer $g(\cdot) = \exp(\cdot)$. Dessa forma, o modelo (2.4) pode ser reescrito como

$$h(t|\mathbf{x}) = h_0(t)\exp(\boldsymbol{\beta}^\top \mathbf{x}). \tag{2.5}$$

Quando o componente $h_0(\cdot)$ não é especificado, o modelo (2.5) é então denominado modelo de riscos proporcionais de Cox. O termo $\exp(\cdot)$ é interpretado como um risco relativo, uma vez que apresenta a razão entre os riscos de ocorrência do evento de interesse para um indivíduo com vetor de covariáveis $\mathbf{x}$ e um indivíduo para o qual $\mathbf{x} = \mathbf{0}$.

Nos modelos de regressão da forma (2.5) as covariáveis têm efeito multiplicativo sobre a função de risco, o que é bastante razoável na maioria das vezes. Entretanto, em algumas situações, o efeito das covariáveis é expresso de forma aditiva na função de risco. Nestes casos, os modelos são denominados modelos aditivos. Tais modelos não são muito comuns, uma vez que a estimação dos parâmetros é, de certa forma, complicada e geralmente feita por meio de métodos não-paramétricos (Fogo, 2007).

O modelo de riscos proporcionais de Cox também pode ser tratado parametricamente. Nesse caso, $h_0(\cdot)$ assume uma distribuição de probabilidade paramétrica. Se uma distribuição de probabilidade adotada é válida, como resultado direto dessa suposição, as estatísticas e inferências para o modelo considerado, obtidas a partir de uma análise completamente paramétrica, serão muito mais precisas. Do ponto de vista paramétrico, seja $f(t)$ a função densidade de probabilidade da variável aleatória associada aos tempos de sobrevivência. Se não há observações censuradas, então a função de verossimilhança para n observações é dada por $\prod_{i=1}^{n} f(t_i)$. Suponha agora que os dados estão sujeitos a censura à direita e considere uma variável indicadora de falha, c_i, em que $c_i = 1$ se o i-ésimo tempo de sobrevivência, $t_i, i = 1, 2, \ldots, n$

é completamente observado, e $c_i = 0$ se o tempo de sobrevivência é censurado. Nesse caso, a função de verossimilhança é dada por

$$\prod_{i=1}^{n} \{f(t_i)\}^{c_i} \{S(t_i)\}^{1-c_i}. \tag{2.6}$$

2.2 Análise de sobrevivência com fração de cura

Em estudos tradicionais envolvendo dados de sobrevivência, assume-se que cada indivíduo na amostra está suscetível ao evento de interesse e, nestes casos, indivíduos que não apresentam o evento de interesse até certo momento são considerados censurados. Assim, estes estudos não levam em consideração o caso em que os indivíduos podem não sofrer tal evento. Em algumas situações porém, pode haver um número de indivíduos para os quais o evento de interesse não se manifestará, independentemente do tempo durante o qual eles foram acompanhados. Por exemplo, em estudos clínicos a população pode responder favoravelmente a uma determinada intervenção, tal como um tratamento, e deixar de ser suscetível ao evento, sendo considerada curada ou imune a tal evento. Modelos que consideram que uma parte da população pode ser ou se tornar não suscetível a certo evento de interesse são denominados modelos com fração de cura ou modelos de longa duração. O termo "longa duração" refere-se aos indivíduos não susceptíveis ao evento de interesse. Na área médica, é comum utilizar o termo "curado" para se referir à parte da população que não está mais em risco.

Modelos de longa duração foram primeiramente abordados por Boag (1949) e Berkson & Gage (1952), que supõem a existência de uma possível causa interferindo para a ocorrência do evento, e que esta causa se manifesta ou não segundo uma probabilidade a ser estimada. Estes modelos ficaram conhecidos na literatura como modelos de mistura padrão. O mecanismo

probabilístico é descrito como segue. Seja T a variável aleatória que representa o tempo até a ocorrência do evento de interesse. Associado ao indivíduo i tem-se uma variável aleatória Bernoulli, Z_i, com probabilidade de sucesso p, tal que Z_i assume o valor 1 se o indivíduo i é suscetível ao evento, e o valor 0 correspondendo a um indivíduo imune. Dessa forma, p representa a proporção de indivíduos susceptíveis na população. Na realidade, não se sabe se um indivíduo é ou não imune, assim Z_i não é observado. Os indivíduos susceptíveis em algum momento apresentarão o evento de interesse, com função de sobrevivência $S(t)$ própria, ou seja, $S(\infty) = 0$. Já os indivíduos com $Z_i = 0$ não apresentarão o evento de interesse, ou seja, seu tempo de ocorrência é infinito. Consequentemente, para todo $t \geq 0$, temos

$$\Pr\{T > t \mid Z_i = 1\} = S(t),$$

$$\Pr\{T > t \mid Z_i = 0\} = 1.$$

Estas probabilidades implicam que, considerando uma população em que existe a possibilidade de indivíduos imunes, a probabilidade de o evento ocorrer após o tempo t, para um indivíduo qualquer é dada por

$$\begin{aligned} S_{\text{pop}}(t) &= \Pr\{T > t\} \\ &= \Pr\{T > t \mid Z_i = 0\}\Pr\{Z_i = 0\} + \Pr\{T > t \mid Z_i = 1\}\Pr\{Z_i = 1\} \\ &= 1 - p + pS(t), \end{aligned} \tag{2.7}$$

em que S_{pop} denota a função de sobrevivência da população. A função de sobrevivência (não condicional) em (2.7) é imprópria, uma vez que $S_{\text{pop}}(\infty) = 1 - p$ e corresponde à proporção de indivíduos curados ou imunes.

Um modelo de longa duração alternativo, proposto e investigado por Yakovlev & Tsodikov (1996) e Chen *et al.* (1999), assume que a função de sobrevivência da população é dada por

$$S_{\text{pop}}(t) = \exp\{-\theta F(t)\}, \quad \theta > 0, \tag{2.8}$$

em que $F(t)$ é uma função de distribuição própria. A proporção de indivíduos imunes, neste caso, é dada por $S_{\text{pop}}(\infty) = \exp\{-\theta\}$.

Covariáveis podem facilmente ser associadas à proporção de susceptíveis em (2.7), através de uma estrutura de regressão logística, de modo que

$$p_i = \frac{\exp(\boldsymbol{b}^\top \mathbf{x}_i)}{1 + \exp(\boldsymbol{b}^\top \mathbf{x}_i)}$$

sendo $\boldsymbol{b}$ o vetor de coeficientes de regressão e $\mathbf{x}_i$, $i = 1, \cdots, n$, o vetor de covariáveis. Da mesma forma, o parâmetro θ pode ser definido em termos das covariáveis, de modo que $\theta_i = \exp(\boldsymbol{b}^\top \mathbf{x}_i)$. Diversas pesquisas têm contribuído com esta área, como os estudos apresentados em Tsodikov *et al.* (2003), Ibrahim *et al.* (2005), Yin & Ibrahim (2005), Peng *et al.* (2007), Yu & Peng (2008), Rodrigues *et al.* (2009), Rodrigues *et al.* (2011) e Cancho *et al.* (2012), que propõem modelos mais abrangentes.

2.3 Modelos de fragilidade

Uma parte substancial da literatura sobre análise de dados de sobrevivência diz respeito aos denominados modelos de fragilidade. Esta classe de modelos considera a existência de uma possível associação entre os tempos de sobrevivência dos indivíduos, sendo caracterizada pela introdução de uma variável aleatória não observada, a fragilidade.

Os modelos de fragilidade podem ser utilizados tanto em estudos de sobrevida univariados quanto multivariados. O termo fragilidade, indicando medida de associação, foi introduzido na análise de sobrevivência por Vaupel *et al.* (1979) e vem sendo alvo de diversas pesquisas, dentre estas Oakes (1982), Hougaard (1984, 1986a) e Hougaard (2000). Modelos de fragilidade para dados univariados têm sido muito utilizados para considerar a heterogeneidade entre indivíduos. Por outro lado, modelos de fragilidade para dados multivariados acrescentam ainda que a fragilidade pode ser utilizada para

modelar a correlação intra-indivíduos. Considerações sobre estes modelos podem ser encontradas em Clayton (1978) e Hougaard (1986b, 2000), em que grande parte do desenvolvimento nesta área decorre dos métodos utilizados para modelar a correlação de dados bivariados de sobrevivência com funções de risco arbitrárias, incluindo os modelos de Cox.

O modelo de fragilidade clássico assume um modelo de risco proporcional condicionado ao efeito aleatório (fragilidade). Este efeito aleatório, considerado em geral uma variável aleatória contínua não-negativa, é incorporado na função de risco como um fator multiplicativo. O fato desse efeito atuar de forma multiplicativa na função de risco é, a princípio, arbitrário, mas tem sido utilizado na maioria dos trabalhos nessa área (Giolo, 2003). Mais especificamente, o efeito de uma fragilidade Z individual é alterar a função de risco basal $h_0(t)$ para $Zh_0(t)$. A função de sobrevivência correspondente, condicionada a Z, é então escrita como

$$S(t \mid Z) = \Pr\{T > t \mid Z\} = \exp\left\{-Z\int_0^t h_0(u)du\right\} = S_0(t)^Z, \qquad (2.9)$$

em que $S_0(t)$ é a função de sobrevivência basal. A função de sobrevivência incondicional, $S(t)$, pode ser obtida integrando (2.9) com respeito à distribuição de Z, uma vez que a distribuição da fragilidade tenha sido especificada. As distribuições de fragilidade frequentemente utilizadas incluem a Gama (Vaupel *et al.*, 1979), Gaussiana Inversa (Hougaard, 1984), Log-normal (Santos *et al.*, 1995) e a família de distribuições estáveis positivas (Hougaard, 1986b; Duchateau & Janssen, 2008). Estas distribuições são convenientes devido à flexibidade e facilidade algébrica em derivar formas fechadas das funções de sobrevivência, densidade e de risco.

Por outro lado, as distribuições de fragilidade contínuas não permitem a possibilidade de risco zero. Sendo assim, existem situações em que uma distribuição discreta pode ser apropriada. Isto pode representar, por exemplo, o número desconhecido de defeitos em uma unidade em teste, ou o número

desconhecido de causas que levam à exposição a determinado dano (Caroni *et al.*, 2010). Ainda, fragilidade zero corresponde a um modelo contendo uma proporção de unidades que nunca falham, que em um contexto médico representa os indivíduos imunes ou curados. Recentemente, alguns autores têm desenvolvido modelos de fragilidade discreta, dentre eles Wienke (2010) aborda o modelo de fragilidade binário e Caroni *et al.* (2010); Ata & Özel (2013) consideram modelos de riscos proporcionais paramétricos com fragilidade discreta permitindo distribuições como a Geométrica, Poisson, Binomial Negativa e outras distribuições discretas para a fragilidade.

Nesse contexto, a fragilidade Z assume valores inteiros não-negativos, ou seja, Z tem distribuição discreta com suporte $\{0, 1, 2, \cdots\}$ ao invés de uma distribuição contínua com suporte em $(0, \infty)$. Seja a distribuição de probabilidade de Z especificada por $\Pr\{Z = z\} = p_z$ para $z = 0, 1, 2, \cdots$. Então, assumindo um modelo de riscos proporcionais, a função de sobrevivência incondicional de T é dada por

$$S(t) = \mathrm{E}\{S_0(t)^Z\} = \sum_{z=0}^{\infty} p_z S_0(t)^z = G_Z\{S_0(t)\}, \tag{2.10}$$

em que G_Z é a função geradora de probabilidade de Z. O caso $z = 0$ implica $\Pr\{T > t \mid Z = 0\} = 1$ para todo t. Distribuições de fragilidade que permitem $p_0 > 0$ podem gerar unidades com fragilidade zero. Para essas unidades, o modelo de riscos proporcionais comporta risco zero, isto é, $S_0(t)^0 = 1$ para todo t, podendo, assim, ser utilizado para descrever as unidades que nunca irão falhar (sobreviventes a longo prazo).

2.4 Eventos recorrentes

Em diversas áreas de estudo encontramos situações em que alguns eventos de interesse não são terminais, isto é, não se encerram no instante de sua ocorrência, podendo ocorrer mais de uma vez para o mesmo indivíduo durante

o perído de observação. Tais eventos são conhecidos na literatura como eventos recorrentes.

Em ciência e tecnologia, o interesse muitas vezes está centrado em estudar processos que geram eventos repetidamente ao longo do tempo. Estes processos são referidos como processos de eventos recorrentes. Esses tipos de processos, segundo Cook & Lawless (2007), surgem frequentemente em estudos longitudinais envolvendo múltiplos sujeitos e podem ser observados em diversas áreas do conhecimento. Exemplos de eventos recorrentes incluem a ocorrência de ataques epiléticos em estudos de neurologia, episódios de pneumonia em pacientes com síndrome de imunodeficiência humana, tumores, ocorrência de cáries ou inflamações em estudos de saúde oral, ataques de asma, episódios de hipoglicemia em diabéticos e gripes, entre outros. Eventos recorrentes também são bastante comuns em estudos de engenharia e confiabilidade quando se observa a ocorrência de falhas ou avarias em um equipamento ou máquina durante um estudo de sua vida útil (Tomazella, 2003).

Os objetivos em estudos que envolvem dados de eventos recorrentes, em geral, incluem caracterizar e descrever o processo de recorrência para os indivíduos, bem como comparar tratamentos com base no tempo até a ocorrência de cada evento. Também é de interesse identificar e explicar a natureza da variação entre os indivíduos, além de determinar a relação entre covariáveis e outros fatores, que muitas vezes não podem ser observados ou medidos, para a ocorrência do evento.

Para a análise de eventos recorrentes, existem essencialmente duas possíves escalas de tempo que podem ser de interesse: (i) o tempo total (ou tempo de calendário), medido a partir do início do acompanhamento até a ocorrência de todos os eventos, e (ii) o tempo entre eventos (ou tempo entre sucessivas ocorrências, tempo entre falhas, *gap time*) (Kelly & Lim, 2000). Dependendo da escala de tempo selecionada, a interpretação da evolução temporal se torna

diferente tanto em modelos paramétricos quanto em modelos semiparamétricos. Entretanto, algumas vezes é útil definir mais de uma escala de tempo. Modelos que incorporam ambas escalas de tempo podem fornecer uma visão mais ampla sobre o que é mais apropriado para um determinado problema. Além da escala temporal, dados de eventos recorrentes também envolvem a escolha de origem, que requer alguns cuidados quando há diversos indivíduos sob estudo. Nesse caso, é necessário que o pesquisador adote uma origem que seja coerente para todos os indivíduos, facilitando a análise e interpretação dos dados (Gouvêa, 2010). Em estudos clínicos, por exemplo, Cook & Lawless (2007) discutem que é habitual considerar o tempo de início do tratamento como o tempo de origem. Isso é bastante razoável, pois geralmente o interesse é fazer comparações entre os tratamentos. Os referidos autores ainda ressaltam que a análise de eventos recorrentes pode ter vários objetivos, levando à formulação de diferentes métodos e modelos para este tipo de dados. Dessa forma, é importante decidir se é mais adequado desenvolver modelos baseados em tempo total ou tempo entre eventos. De acordo com Gouvêa (2010), embora esse fato possa ser visto como uma decisão na especificação do modelo, afeta a análise e interpretação dos resultados.

A modelagem de dados de eventos recorrentes pode ser considerada de diversas forma. As abordagens que são frequentemente mais utilizadas para modelagem e análise estatística desses dados são aquelas baseadas nos conceitos de função de intensidade e processos de contagem (Tomazella, 2003; Cook & Lawless, 2007; Aalen *et al.*, 2008).

2.4.1 Notação e conceitos básicos

Segundo Carvalho *et al.* (2005), uma característica importante dos dados de eventos recorrentes é a dependência temporal ou estocástica, permitindo assim a modelagem dos eventos através de métodos de sobrevivência baseados na teoria de processos de contagem. Modelos estatísticos baseados nesta

teoria para analisar dados de eventos recorrentes foram originalmente introduzidos por Aalen (1980). Aalen (1978) apresenta os elementos da teoria de processos de contagem, utilizados no âmbito da inferência estatística, e o seu papel fundamental na base matemática da análise de sobrevivência. Uma descrição bastante detalhada da teoria de processos de contagem ainda pode ser encontrada em Fleming & Harrington (1991) e Andersen *et al.* (1993).

Considere um processo de eventos recorrentes iniciando-se no tempo $t_0 = 0$ e sejam $0 < T_1 < T_2 < \ldots$ os tempos de falha contínuos (tempos dos eventos). Estes tempos de falha podem ser vistos como a realização de um processo pontual simples na reta real. Os tempos de falha geram um processo de contagem $\{N(t) : t \geq 0\}$, em que $N(t)$ registra o número de ocorrências de um evento recorrente no intervalo $[0, t]$. Uma outra representação do mesmo processo pode ser dada pelos tempos entre eventos, definidos por $Y_j = T_j - T_{j-1},\ j \geq 1$ e $T_0 = 0$. A Figura 2.1 apresenta uma realização de um processo de eventos recorrentes em termos de seu processo de contagem.

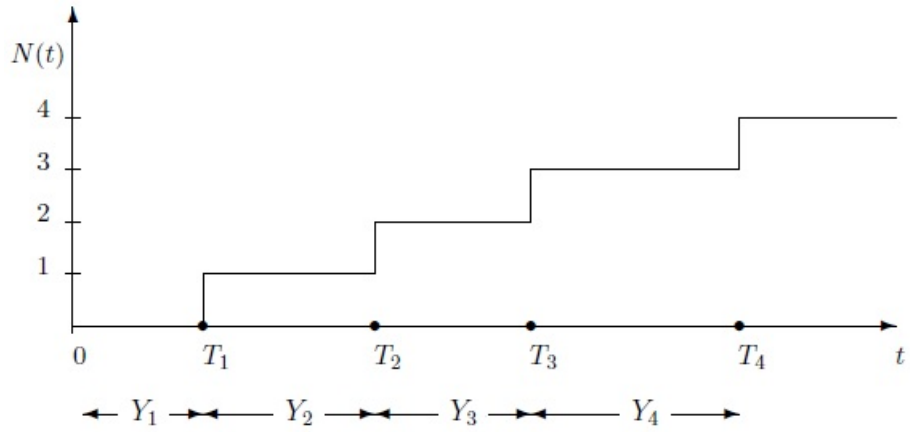


Figura 2.1: Representação do processo de contagem com dados de eventos recorrentes.

Assumindo-se que as falhas em um processo de eventos recorrentes são equivalentemente definidas pelos processos $\{N(t)\}_{t\geq 0}$, $\{T_j\}_{j\geq 1}$ ou $\{Y_j\}_{j\geq 1}$,

e que não ocorrem falhas simultâneas, a estrutura probabilística de tais processos pode ser descrita em termos de um processo de intensidade, cuja função de intensidade, também denominada de intensidade condicional, é definida como

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr\{N(t+dt) - N(t) = 1 \mid \mathcal{F}_{t^-}\}}{dt}, \qquad \forall t \geq 0 \qquad (2.11)$$

em que $\mathcal{F}_{t^-}$ é a história ou filtragem gerada pelo processo de contagem, que consiste do conhecimento acerca do que aconteceu aos indivíduos até um instante imediatamente anterior a t. Formalmente, o processo de intensidade pode ser reescrito como $\lambda(t)dt = \mathrm{E}[dN(t)|\mathcal{F}_{t^-}]$, em que $dN(t) = N(t+dt^-) - N(t^-)$ representa o número de eventos em um curto intervalo de tempo $(t, t+dt]$ (Andersen *et al.*, 1993; Aalen *et al.*, 2008). O incremento $dN(t)$ é tal que $dN(t) = 1$ se uma falha ocorreu em t, e $dN(t) = 0$ se não ocorreu falha em t.

É importante comparar as definições da função de risco em (2.2) e da função de intensidade em (2.11). A função de risco $h(t)$ é a probabilidade de que o evento de interesse ocorra apenas uma vez no intervalo de tempo $(t, t+dt]$, condicionada à sobrevivência no instante t, dividida pelo tamanho do intervalo. A função de intensidade $\lambda(t)$ é a probabilidade incondicional de ocorrência do evento (não necessariamente o primeiro) no intervalo $(t, t+dt]$, dividida pelo tamanho do intervalo.

Outra característica importante é a função de média do processo de contagem $N(t)$, denotada por $\Lambda(t)$, a qual é também um processo estocástico com a propriedade $\Lambda(t) = \mathrm{E}[N(t)|\mathcal{F}_{t^-}]$.

Muitos autores, incluindo Prentice *et al.* (1981) e Andersen & Gill (1982), consideram os eventos recorrentes de um particular indivíduo como a realização de um processo de contagem e modelam a função de intensidade do processo de recorrência. Diversos modelos baseados em intensidade podem ser estabelecidos a partir de (2.11). No entanto, podemos dividí-los em dois tipos

fundamentais: (i) aqueles que são caracterizados por meio das propriedades dos tempos de ocorrência do evento, e (ii) aqueles caracterizados por meio da distribuição dos tempos entre eventos. Sendo assim, métodos para análise de dados de eventos recorrentes geralmente baseiam-se em contagem e função taxa ou análise de tempos entre eventos.

Normalmente, um modelo de intensidade depende da história do processo. Uma exceção para isso são os processos de Poisson (Finkelstein, 2008), em que a história do processo até um tempo $t^- < t$ não afeta a probabilidade de ocorrência do evento, e na ausência de covariáveis, o único fator que determina a intensidade é o tempo t. Estes processos são considerados modelos marginais e formam uma classe de processos de contagem muito importante em aplicações práticas. O processo de Poisson é definido como segue.

Definição 2.4.1. Um processo de contagem $\{N(t) : t \geq 0\}$ é dito ser um processo de Poisson se

1. $N(0) = 0$;

2. $N(t)$ tem incrementos independentes, isto é, se $(a, b]$ e $(c, d]$ são intervalos disjuntos então $N(a, b)$ e $N(c, d)$ são variáveis aleatórias independentes;

3. O processo tem função de intensidade

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr\{N(t + dt) - N(t) = 1\}}{dt};$$

4. O processo é regular, ou seja,

$$\lim_{dt \to 0} \frac{\Pr\{N(t + dt) - N(t) \geq 2\}}{dt} = 0.$$

Para processos de Poisson a função de intensidade $\lambda(t)$ é também a função de taxa (ou taxa de ocorrência de falhas, *rocof*). Tais processos podem ser classificados em homogêneos e não homogêneos, dependendo da sua função de intensidade. Como resultado da Definição 2.4.1, a variável aleatória $N(t)$ tem

distribuição de Poisson com média $\int_0^t \lambda(u)du$. Um dos modelos amplamente empregados para a análise de eventos recorrentes é o processo de Poisson não homogêneo (PPNH) (vide por exemplo, Lawless (1987); Rigdon & Basu (2000); Lawless (2003)). Neste modelo a função de intensidade $\lambda(t)$ varia com o tempo e, o número médio de recorrências até o tempo t é dado por $\Lambda(t) = \int_0^t \lambda(u)du$, em que $\Lambda(t)$ é a função de média do processo também conhecida como função de intensidade acumulada. Um caso particular do PPNH é o processo de Poisson homogêneo (PPH), em que a função de intensidade não depende de t, isto é, $\lambda(t) = \lambda$ é uma função constante. Neste caso, $\Lambda(t) = \lambda t$ e os tempos entre eventos são variáveis aleatórias independentes e identicamente distribuídas com distribuição Exponencial de média $1/\lambda$.

Um outro conceito importante está ligado à ideia de um conjunto de elementos sob risco, ou seja, quando um certo número de indivíduos está em risco para a ocorrência de determinado evento. Sendo assim, de acordo com a teoria de processos de contagem, para cada indivíduo podemos então observar dois processos; o processo de contagem $N(t)$, e o processo indicador de risco $R(t)$, em que $R(t) = 1$ se o indivíduo está em observação e, portanto, sujeito ao risco do evento no instante t, e $R(t) = 0$ caso contrário. A função $R(t)$ é usada para denotar quais indivíduos fornecem informações sobre a ocorrência do evento em um dado tempo. Dessa forma, a função de intensidade associada ao processo pode ser expressa em termos de $R(t)$. Nesse caso, o incremento $dN(t)$ no intervalo $(t, t + dt]$, condicionado a $\mathcal{F}_{t^-}$, pode ser visto como tendo uma distribuição de Bernoulli com

$$\Pr\{dN(t) = 1|\mathcal{F}_{t^-}\} = R(t)\lambda(t)dt.$$

Assim, pelos resultados de Jacod (1975), pode-se escrever o processo de verossimilhança no tempo t para um particular indivíduo, em termos de um parâmetro θ, como

$$\mathcal{L}(\theta; t) = \left[\prod_{t\geq 0} \{R(t)\lambda(t;\theta)\}^{dN(t)}\right] \times \exp\left\{-\int_0^\infty R(u)\lambda(u;\theta)du\right\}, \quad (2.12)$$

em que $\prod$ denota o produto integral. Para detalhes sobre produto integral vide Gill & Johansen (1990); Andersen *et al.* (1993).

Consequentemente, o processo de log-verossimilhança $\{\ell(\theta;t) : t \geq 0\}$ é dado por

$$\ell(\theta;t) = \int_0^\infty \log\{R(t)\lambda(t;\theta)\}dN(t) - \int_0^\infty R(t)\lambda(t;\theta)dt. \tag{2.13}$$

Uma introdução bastante abrangente para análise de dados de eventos recorrentes pode ser encontrada no livro de Cook & Lawless (2007), que discutem uma série de diferentes abordagens neste contexto.

2.4.2 Referencial teórico

Existe uma extensa literatura sobre dados de eventos recorrentes. De acordo com Cook & Lawless (2007), devido aos diversos objetivos da análise de eventos recorrentes, há necessidade da formulação de diferentes métodos e modelos para a análise de tais dados. Nesse contexto, modelos canônicos como o processo de Poisson, que modela o tempo total de estudo, e o processo de renovação, que modela os tempos entre eventos, têm sido amplamente estudados e utilizados. Tais processos têm propriedades simples e de fácil interpretação. No entanto, sua gama de aplicação é limitada e na maioria das situações temos que considerar tanto as extensões destes processos quanto outros modelos alternativos formulados através de funções de intensidade. A maioria dos métodos de regressão existentes para análise de dados de eventos recorrentes assume efeito de covariáveis multiplicativo. Diversos autores têm considerado modelos para eventos recorrentes com base no modelo de riscos proporcionais de Cox. Therneau & Grambsch (2000) distingue três categorias para os modelos: incrementos independentes, marginais e condicionais, sendo que a última inclui os modelos com efeitos aleatórios ou fragilidade.

Andersen *et al.* (1993) apresenta uma descrição completa de métodos para modelos de intensidade, enfatizando os modelos de Markov modulados.

Kalbfleisch & Prentice (2002); Aalen *et al.* (2004); Fosen *et al.* (2006a,b) discutem métodos para análise de dados de eventos recorrentes usando covariáveis dinâmicas. Andersen & Gill (1982) propõem um modelo para eventos recorrentes assumindo incrementos independentes. Este modelo considera que o risco basal é igual em todos os intervalos de tempo analisados, sendo que o indivíduo retorna ao grupo de risco após cada evento. Além disso, para uma função de risco basal paramétrica, o modelo de Andersen-Gill é equivalente ao processo de Poisson não homogêneo. Cox (1972b) introduz os modelos de renovação modulados, enquanto Berman & Turner (1992) e Lawless & Thiagarajah (1996) consideram modelos mais gerais, vistos como extensões dos modelos de Cox (1972b), e que acomodam ambas as escalas de tempo, tempo total e tempo entre eventos. Nesse mesmo contexto, Louzada-Neto (2004, 2008) propõe modelos paramétricos híbridos para dados de eventos recorrentes, admitindo ambas as escalas de tempo além da contagem do número de eventos para cada indivíduo. Estes modelos englobam uma ampla classe de modelos de intensidade, incluindo os processos de Poisson e renovação como casos particulares. Prentice *et al.* (1981) propõem um modelo semiparamétrico para eventos recorrentes que mede o risco condicional de experimentar um evento. O modelo separa a análise em diferentes estratos, assumindo que existe uma dependência entre os tempos de falha de um mesmo indivíduo. O uso de estratos dependentes significa que a função intensidade pode variar de um evento para outro, ao contrário do que ocorre no modelo de Andersen-Gill. O modelo de Prentice *et al.* (1981) pressupõe ainda que um indivíduo só estará em risco de experimentar o m-ésimo evento depois que tenha experimentado o evento $(m-1)$. Alternativamente, utilizando uma abordagem marginal, Wei *et al.* (1989) propõem um modelo estratificado para a análise de tempos de falha recorrentes, em que a ordem neste caso é dada pela enumeração de cada evento. Neste modelo, o indivíduo ao entrar no estudo está simultaneamente em risco de experimentar os m eventos, que podem ocorrer no estudo. Os

autores consideram a modelagem da distribuição marginal de cada tempo de falha, sem modelar explicitamente a correlação entre os eventos observados.

Modelos marginais para análise das funções de taxa e média do processo de contagem para eventos recorrentes também têm sido estudados por diversos autores, dentre estes Lawless & Nadeau (1995); Lin *et al.* (2000); Scheike (2002); Chiang *et al.* (2005); Cook & Lawless (2007); Martinussen & Scheike (2007). Em um outro trabalho (Fredette & Lawless, 2007) são estudados métodos de predição para eventos recorrentes que ocorrem para indivíduos ou unidades em uma mesma população utilizando o processo de Poisson não homogêneo. Ainda, Ghosh (2004) e Sun & Su (2008) propõem modelos de risco acelerados, e Schaubel *et al.* (2006) e Lim & Zhang (2009) consideram modelos de risco aditivos semiparamétricos. Lin *et al.* (1998) e Jin *et al.* (2006) abordam modelos de tempo de falha acelerado semiparamétricos para processos de contagem. Outro trabalho (Zeng & Lin, 2006) propõe uma classe de modelos de transformação semiparamétricos para processos de contagem em geral, a qual incorpora modelos mais flexíveis para eventos recorrentes. Mais recentemente, Crowther & Lambert (2014) apresentam uma estrutura geral para modelos paramétricos incluindo eventos recorrentes, e Dong & Sun (2015) apresentam uma classe de modelos de transformação semiparamétricos para a análise de dados de eventos recorrentes que permite ambos os efeitos de covariáveis, multiplicativo e aditivo, e ainda covariáveis dependentes do tempo.

Para os casos em que o tempo entre os sucessivos eventos é a variável de interesse, a estrutura estocástica dos dados de eventos recorrentes gera mudanças na análise estatística, em relação à análise dos processos de Poisson, requerendo assim uma modelagem apropriada. Nesse contexto, podemos destacar uma variedade de trabalhos. Entre estes, Chen *et al.* (2004) consideram o problema de regressão para as distribuições dos tempos entre eventos usando um modelo de riscos proporcionais tempo-reverso. Sreeja & Sankaran (2007)

apresentam um modelo semiparamétrico para avaliar a relação entre a vida média residual e covariáveis para as distribuições dos tempos entre eventos. Outros modelos como o modelo de tempo de falha acelerado (Strawderman, 2005) e os modelos de riscos aditivos e multiplicativos (Huang & Chen, 2003; Sun *et al.*, 2006; Lim & Zhang, 2011) também são encontrados na literatura para a análise dos tempos entre eventos. Outro trabalho, (Luo & Huang, 2011) demonstra que muitos métodos existentes para a análise dos tempos entre eventos podem ser vistos como métodos de riscos definidos ponderados. Ainda, Sankaran & Anisha (2012) consideram um modelo de riscos aditivos para os tempos entre eventos com múltiplas causas e Zhao & Zhou (2012) apresentam um modelo aditivo semiparamétrico, o qual é derivado de um processo de Poisson não homogêneo. Mais recentemente, Zhu (2014) fornece uma visão geral dos métodos de estimação não paramétricos existentes para a distribuição dos tempos entre eventos recorrentes.

Outra abordagem comum na análise de eventos recorrentes é o uso de modelos de fragilidade, em que uma variável aleatória contínua é introduzida no modelo a fim de explicar a heterogeneidade existente entre os indivíduos. Para eventos recorrentes a fragilidade é geralmente utilizada para modelar a dependência entre os tempos de um mesmo indivíduo. Uma visão geral destes modelos especialmente formulados para dados de recorrência é dada por Oakes (1992); Hougaard (2000); Duchateau *et al.* (2003); Bijwaard *et al.* (2006); Lim *et al.* (2007). Diferentes abordagens no sentido de incorporar dependência e heterogeneidade nos modelos de eventos recorrentes têm sido consideradas por diversos autores, por exemplo Duchateau *et al.* (2003) aplicam diferentes modelos paramétricos e não paramétricos, com e sem fragilidade, para os dados de um estudo referente à episódios recorrentes de asma. Box-Steffensmeier & De Boef (2006) abordam modelos para análise de eventos recorrentes que incluem tanto heterogeneidade quanto dependência de eventos usando um termo de fragilidade. Zeng & Lin (2007) propõem uma classe de modelos

semiparamétricos com efeitos aleatórios para eventos recorrentes e Sankaran & Anisha (2011) apresentam um modelo de fragilidade semiparamétrico com múltiplas causas. Mais recentemente, Liu *et al.* (2014) propõem um modelo de intensidade acelerado com fragilidade para dados de eventos recorrentes, e Somboonsavatdee & Sen (2014) apresentam um modelo paramétrico para a análise de sistemas reparáveis múltiplos com riscos competitivos dependentes, em que a dependência entre os processos recorrentes, específico para cada causa, é modelada por meio de um termo de fragilidade. Ainda, Wang *et al.* (2001) consideram situações em que, além dos eventos recorrentes, existem eventos terminais que impedem a ocorrência dos demais eventos, gerando assim censuras dependentes. Zeng & Lin (2009) propõem uma ampla classe de modelos semiparamétricos com efeitos aleatórios para análise de eventos recorrentes na presença de um evento terminal. Em um mesmo contexto, Bao *et al.* (2013) consideram um modelo que incorpora um termo de fragilidade para modelar conjuntamente eventos recorrentes e eventos terminais com aplicação a um estudo de implante dentário, enquanto Sun & Kang (2013) abordam um modelo de riscos aditivos e multiplicativos para dados de eventos recorrentes na presença de um evento terminal como a morte.

Nos últimos anos, um crescente interesse tem surgido também em modelos para dados de sobrevivência os quais assumem uma proporção de indivíduos altamente susceptíveis a um determinado tipo de evento adverso, e outros indivíduos considerados em risco muito menor. Diversas contribuições nesse contexto têm surgido, com muitas aplicações principalmente na área médica. Assim, se um número significante de indivíduos são curados, tornando-se assim livres de recorrências após um primeiro tratamento, a população é então considerada uma mistura de indivíduos susceptíveis e não susceptíveis. Para modelar dados de eventos recorrentes nesta situação são utilizados, geralmente, os modelos de fragilidade e fração de cura. Price & Manatunga (2001) consideram modelos de fragilidade e fração de cura para analisar a

recorrência de leucemia entre paciente transplantados. Entretanto, nesta abordagem é observado para cada indivíduo apenas o tempo até a primeira recorrência e os efeitos aleatórios explicam, então, a heterogeneidade entre os indivíduos devido a fatores de riscos não observados. Yu (2008) propõe um modelo de mistura com fragilidade e fração de cura para analisar dados de reinternações hospitares. Um algoritmo EM (*Expectation-Maximization*) é usado para estimar os parâmetros e os erros padrão são calculados pelo método *boostratp*. Um outro trabalho (Rondeau *et al.*, 2011) compara modelos de mistura com fragilidade e fração de cura para eventos recorrentes considerando diferentes formas para a taxa de cura. Os autores consideram um modelo que permite uma probabilidade de cura após cada evento. Mais recentemente, Louzada & Cobre (2012) propõem um modelo de escala múltipla de tempo para analisar dados de eventos recorrentes com fração de cura e Xu *et al.* (2014) propõem um modelo de fragilidade e fração de cura para a estimativa do efeito de uma intervenção sobre a probabilidade de cura e o efeito total sobre a taxa de eventos no grupo de indivíduos não curados.

Capítulo 3

Modelo para Tempos entre Eventos Recorrentes (Modelo I)

Dados de eventos recorrentes têm sido recentemente investigados por diversos autores baseados tanto na função de taxa marginal do processo de recorrência (Cook & Lawless, 2002; Lin *et al.*, 2000; Wang *et al.*, 2001), quanto nos tempos de ocorrência dos eventos (McDonald & Rosina, 2001), ou tempos entre os sucessivos eventos (Huang & Chen, 2003; Schaubel & Cai, 2004a).

Considerável atenção também tem sido dada aos modelos para tempos entre eventos, incluindo a modelagem de distribuições multivariadas dos tempos entre sucessivas recorrências do evento (Lin *et al.*, 1999), modelos de riscos aditivos e multiplicativos baseados em processo de renovação (Huang & Chen, 2003; Sun *et al.*, 2006), modelos de riscos proporcionais sem especificação de uma estrutura de dependência entre os indivíduos (Schaubel & Cai, 2004b), entre outros. Estes modelos têm sido construídos com base na suposição de risco aditivos ou multiplicativos. Essa suposição, no entanto, não reflete adequadamente muitas situações práticas. Dessa forma, segundo Wang *et al.* (2001), os modelos de taxa marginais muitas vezes são preferidos devido ao fato de fornecerem interpretações práticas mais diretas para a identificação

de riscos do que os modelos para tempos entre eventos como acima citados.

Nesse contexto, propomos um modelo de taxa marginal para analisar tempos entre eventos recorrentes, baseado na ideia de Zhao & Zhou (2012), no qual o processo de recorrência segue um modelo de Poisson não homogêneo e a função de taxa associada é caracterizada por uma estrutura multiplicativa. A função de taxa pode assumir uma forma paramétrica ou semiparamétrica, se a função basal for especificada por um vetor de parâmetros ou arbitrária, respectivamente. Neste trabalho, assume-se uma distribuição Weibull para a função basal, devido a sua flexibilidade e versatilidade em acomodar diversas situações práticas.

Contudo, o modelo proposto é atrativo e tem vantagens tanto de uma interpretação direta da função de taxa marginal para fins práticos, quanto de uma inferência estatística eficaz a partir dos tempos entre eventos recorrentes. A metodologia desenvolvida é avaliada através de conjuntos de dados simulados e a análise envolve um conjunto de dados reais.

3.1 Formulação geral do modelo

Suponha que um indivíduo sob observação durante um período de tempo $[0, \tau]$ pode experimentar consecutivas recorrências de um mesmo tipo de evento nos tempos $0 < T_1 < T_2 < \cdots < T_j < \cdots$, $j = 1, 2, \ldots$, medidos a partir do início do estudo. Os tempos entre eventos são então definidos como $Y_j = T_j - T_{j-1}$, tempo entre o $(j-1)$-ésimo e j-ésimo evento, para $j = 1, 2, \cdots$ e $T_0 = 0$.

Para uma representação explícita do modelo proposto, considere t como o tempo de calendário e y como o tempo de interesse (tempo entre eventos). Com isso, neste trabalho assume-se um modelo multiplicativo, baseado em (2.5), para a função de taxa do processo de recorrência dado por

$$\lambda(y|t, \mathbf{x}) = \lambda_0(y + t) \exp(\boldsymbol{\beta}^\top \mathbf{x}), \tag{3.1}$$

em que $\lambda(y|t,\mathbf{x})$ é a função de taxa do processo de recorrência individual até o tempo $y+t$ com vetor de covariáveis $\mathbf{x}$, λ_0 é uma função basal contínua que descreve o comportamento geral do indivíduo ao longo do tempo e $\boldsymbol{\beta}$ é o vetor de coeficientes de regressão associado à $\mathbf{x}$. O processo de recorrência $N(y+t)$ de um indivíduo arbitrário com vetor de covariáveis $\mathbf{x}$ é então assumido ser um PPNH com função de taxa associada dada por (3.1).

Seja a função de taxa acumulada sobre o intervalo $(t, t+y]$ expressa por

$$\Lambda(t,y) = \Lambda(t+y) - \Lambda(t) = \int_0^y \lambda(u|t,\mathbf{x})du, \tag{3.2}$$

com $\Lambda_0(t,y) = \int_0^y \lambda_0(u+t)du$ a função de taxa basal acumulada sobre $(t, t+y]$. Então, sob a suposição de um PPNH com função de taxa (3.1), e $\mathrm{E}[N(t)] = \Lambda(t)$, obtemos

$$\begin{aligned} \mathrm{E}[N(t,t+y)] &= \mathrm{E}[N(t+y) - N(t)] = \Lambda(t+y) - \Lambda(t) = \Lambda(t,y) \\ &= \int_0^y \lambda(u|t,\mathbf{x})du = \int_0^y \lambda_0(u+t)\exp(\boldsymbol{\beta}^\top \mathbf{x})du \\ &= \Lambda_0(t,y)\exp(\boldsymbol{\beta}^\top \mathbf{x}). \end{aligned} \tag{3.3}$$

O conhecimento da função de intensidade de um processo de recorrência nos permite obter características do processo que são frequentemente úteis, em particular, as distribuições condicionais dos tempos entre eventos. Dessa forma, sendo a função de taxa em (3.1) uma função determinística de tempo e integravél sobre o intervalo $(t, t+y]$, pode-se obter a função de sobrevivência de Y_j, condicionada às recorrências anteriores, como segue.

Para um indivíduo arbitrário, seja $T_j = Y_1 + \cdots + Y_j$ o tempo de ocorrência do j-ésimo evento. Assim, pela propriedade de incrementos independentes do processo de Poisson, e considerando a expressão em (3.3), segue que o modelo

(3.1) corresponde a uma função de sobrevivência da forma

$$\begin{aligned} S(y|t) &= \Pr(Y_j > y | T_{j-1} = t) \\ &= \Pr\{N(t+y) - N(t) = 0 | N(t) = j-1\} \\ &= \Pr\{N(t+y) - N(t) = 0\} = \exp\{-\mathrm{E}[N(t+y) - N(t)]\} \\ &= \exp\{-\Lambda(t,y)\} = \exp\{-\Lambda_0(t,y)\}^{\exp(\boldsymbol{\beta}^\top \mathbf{x})}. \end{aligned} \tag{3.4}$$

Além disso, segue de (3.4) que condicionado a $T_{j-1} = t$, o j-ésimo tempo entre eventos Y_j, com função de taxa expressa por (3.1), tem função densidade dada por

$$f_{Y_j}(y|t) = \lambda(y|t, \mathbf{x}) \exp\{-\Lambda(t,y)\}. \tag{3.5}$$

Os seguintes fatos são decorrentes do modelo proposto:

(i) Os tempos entre eventos $\{Y_j : j = 1, 2, \cdots\}$ de um indivíduo são condicionalmente independentes, dado o tempo da recorrência anterior T_{j-1} e o vetor de covariáveis $\mathbf{x}$.

(ii) Condicionado a T_{j-1} e $\mathbf{x}$, o tempo até o primeiro evento Y_1 tem uma distribuição diferente dos demais tempos entre eventos $Y_2, Y_3, \cdots, Y_j, \cdots$.

3.1.1 Construção da função de verossimilhança

Considere agora que n indivíduos são observados independentemente. Para cada indivíduo, seja M_i o número total de recorrências no período em estudo, $i = 1, 2, \cdots, n$, tal que $\Pr\{M_i < \infty\} = 1$. Seja $Y_{i1} = T_{i1},\ Y_{i2} = T_{i2} - T_{i1}, \cdots, Y_{i,M_i} = T_{i,M_i} - T_{i,M_i-1}$ os tempos entre eventos, e $\mathbf{x}_{ij} = (x_{1ij}, \cdots, x_{pij})$ um vetor de p covariáveis externas (fixas ou variando com j) para o indivíduo i.

O tempo de acompanhamento de um indivíduo sob estudo geralmente está sujeito a um tempo de censura C_i, tal como um tempo de falha ou tempo de abandono, o qual é não informativo sobre os tempos de recorrência $(T_1, T_2, \ldots)$

e que satisfaz a condição de censura independente (Andersen *et al.*, 1993, p.139). Dessa forma, como discutido em Aalen & Husebye (1991) e Wang & Chang (1999), o tempo até o primeiro evento quando $M_i = 1$ e o último tempo entre eventos quando $M_i > 1$ são geralmente censurados por C_i. Isto leva a uma dependência induzida entre o tempo de censura e o último tempo entre eventos quando $M_i > 1$, apesar da independência entre C_i e $N_i(\cdot)$, uma vez que um valor maior de Y_{ij} está associado a uma alta probabilidade de $Y_{i,j+1}$ ser censurado. Nesse caso, Wang & Chang (1999), Huang & Chen (2003) e Sun *et al.* (2006) propõem remover o último tempo entre eventos quando $M_i > 1$.

Sob a suposição de independência entre o processo de recorrência $N_i(\cdot)$ e o tempo de censura C_i, para indexar a dependência induzida quando $M_i > 1$ procedemos como segue. Para o indivíduo i, o tempo de acompanhamento é dado por $\tau_i = \min(\tau, C_i)$, em que τ é o tempo de acompanhamento máximo fixado no planejamento do estudo (tempo final de estudo). As observações do processo de recorrência para esse indivíduo são dadas por $\{Y_{ij} : j = 1, 2, \cdots\}$, as quais satisfazem

$$\sum_{j=1}^{M_i-1} Y_{ij} < \tau_i \qquad \text{e} \qquad \sum_{j=1}^{M_i} Y_{ij} \geq \tau_i.$$

Os dados disponíveis são $\{Y_{i1}, \cdots, Y_{i,M_i-1}, \tau_i\}$, em que os primeiros $M_i - 1$ tempos entre eventos são observados, enquanto Y_{i,M_i} é censurado por $Y^+_{i,M_i} = \tau_i - \sum_{j=1}^{M_i-1} Y_{ij}$.

Quando $M_i > 1$, segue de (2.12) que a contribuição dos primeiros $M_i - 1$ tempos entre eventos $\{Y_{i1}, \cdots, Y_{i,M_i-1}\}$, os quais são completamente observados, para a função de verossimilhança é dada por

$$\prod_{j=1}^{M_i-1} \lambda_i(y_{ij}|t_{i,j-1}, \mathbf{x}_{ij}) \exp\left\{-\int_0^\infty R_i(u)\lambda_i(u|t, \mathbf{x})du\right\}, \tag{3.6}$$

em que $R_i(\cdot)$ é o processo indicador de risco para o indivíduo i. Note que como $N_i(\cdot)$ é uma função degrau com saltos de tamanho um, o primeiro termo

na expressão dada por (2.12) se reduz a um produto finito sobre os instantes nos quais ocorrem saltos no processo de contagem ($dN_i(\cdot) = 1$), obtendo-se assim a expressão em (3.6).

Para cada i, defina $d_i = I(M_i > 1)$ e $\tilde{Y}_{ij} = \min(Y_{ij}, \tau_i - T_{i,j-1})$, em que $I(\cdot)$ denota a função indicadora. Assim, considerando os dados observados $\mathcal{D}_i = \{\tilde{y}_{ij}, t_{ij}, \mathbf{x}_{ij}, \tau_i, M_i : i = 1, \cdots, n;\ j = 1, \cdots, M_i\}$, a função de verossimilhança conjunta para todas as observações (completas e incompletas) do indivíduo i é expressa como

$$\mathcal{L}_i(\boldsymbol{\theta}|\mathcal{D}_i) = \left(\prod_{j=1}^{M_i - 1} \lambda_i(\tilde{y}_{ij}|t_{i,j-1}, \mathbf{x}_{ij}) \right)^{d_i} \exp\left\{ - \int_0^{\infty} R_i(u) \lambda_i(u|t, \mathbf{x}) du \right\} \quad (3.7)$$

em que $\boldsymbol{\theta}$ denota o vetor de parâmetros de interesse. Ainda, a integral no segundo termo da expressão (3.7) pode ser escrita como

$$\int_0^{\infty} R_i(u) \lambda_i(u|t, \mathbf{x}) du = \sum_{j=1}^{M_i} \int_0^{y_{ij}} \lambda_i(u|t_{i,j-1}, \mathbf{x}_{ij}) du. \quad (3.8)$$

Portanto, com a informação dos n indivíduos a função de verossimilhança, obtida a partir de (3.7) e (3.8), é expressa como

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{n} \left[\left(\prod_{j=1}^{M_i - 1} \lambda_i(\tilde{y}_{ij}|t_{i,j-1}, \mathbf{x}_{ij}) \right)^{d_i} \exp\left\{ - \sum_{j=1}^{M_i} \int_0^{\tilde{y}_{ij}} \lambda_i(u|t_{i,j-1}, \mathbf{x}_{ij}) du \right\} \right], \quad (3.9)$$

em que $\mathcal{D} = (\mathcal{D}_1, \cdots, \mathcal{D}_n)$ é o conjunto de dados observado.

3.2 Inferência

Nesta seção apresentamos uma abordagem clássica para o modelo proposto na seção anterior. O procedimento de estimação para os parâmetros do modelo é realizado considerando-se o método de máxima verossimilhança. Intervalos de confiança são construídos para os parâmetros do modelo. Testes de hipóteses baseados em verossimilhança, tais como o teste da razão de

verossimilhanças e o teste do escore, bem como uma ferramenta gráfica de diagnóstico são considerados a fim de testar a adequabilidade e avaliar a qualidade do ajuste do modelo.

3.2.1 Estimação pelo método de máxima verossimilhança

Para a construção das equações de verossimilhança e estimação dos parâmetros do modelo, assumimos uma distribuição Weibull para a função de taxa basal λ_0 em (3.1),

$$\lambda_0(y+t) = \alpha\delta(y+t)^{\delta-1}, \tag{3.10}$$

em que $\alpha,\ \delta > 0$ são os parâmetros de escala e forma, respectivamente. O modelo de Weibull é bastante explorado e utilizado devido a sua flexibilidade em acomodar diversas situações práticas. Quando $\delta > 1$, a função de taxa do processo é crescente, e quando $\delta < 1$, a função de taxa é decrescente. Há ainda o caso em que $\delta = 1$, em que a função de taxa é constante no tempo, obtendo-se o PPH como caso particular.

Dessa forma, segue de (3.1) e (3.10) que a função de taxa do modelo para tempos entre eventos associada ao processo de recorrência do i-ésimo indivíduo é dada por

$$\lambda_i(y_{ij}|t_{i,j-1}, \mathbf{x}_{ij}) = \delta(y_{ij} + t_{i,j-1})^{\delta-1}\exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij}), \qquad \delta > 0, \tag{3.11}$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^\top$ e $\mathbf{x}_{ij} = (1, x_{1ij}, \cdots, x_{pij})$. Em termos de covariável zero, ou seja, $\mathbf{x}_{ij} = (1, 0, \cdots, 0)$ tem-se $\alpha = \mathrm{e}^{\beta_0}$.

Considerando a informação referente aos n indivíduos, a função de log-verossimilhança baseada nos dados observados é dada por $\ell(\boldsymbol{\theta}|\mathcal{D}) = \log\{\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})\}$, com $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ dada por (3.9). O vetor de parâmetros de interesse é dado por $\boldsymbol{\theta} = (\delta, \beta_0, \cdots, \beta_p)$. Dessa forma, considerando a função de taxa em (3.11), a

função de log-verossimilhança associada ao modelo proposto é expressa na forma

$$\ell(\boldsymbol{\theta}|\mathcal{D}) = M\log(\delta) + \sum_{i=1}^{n} d_i \left\{ (\delta-1) \sum_{j=1}^{M_i-1} \log(\tilde{y}_{ij} + t_{i,j-1}) + \sum_{j=1}^{M_i-1} \boldsymbol{\beta}^\top \mathbf{x}_{ij} \right\} - \sum_{i=1}^{n} \sum_{j=1}^{M_i} \mathcal{K}_{ij}(\delta) \exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij}), \tag{3.12}$$

em que $M = \sum_{i=1}^{n} d_i(M_i - 1)$ é o número total de eventos observados, $d_i = I(M_i > 1)$ e $\mathcal{K}_{ij}(\delta) = (\tilde{y}_{ij} + t_{i,j-1})^\delta - t_{i,j-1}^\delta$, para $i = 1, \cdots, n$ e $j = 1, \cdots, M_i$.

Com a primeira derivada da função de log-versossimilhança $\ell(\boldsymbol{\theta}|\mathcal{D})$ em relação a cada parâmetro do modelo, definimos a função (ou vetor) escore $\mathbf{U}(\boldsymbol{\theta})$, cujos elementos são dados por

$$U_\delta(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial \delta} = \frac{M}{\delta} + \sum_{i=1}^{n} \left\{ d_i \sum_{j=1}^{M_i-1} \log(\tilde{y}_{ij} + t_{i,j-1}) - \sum_{j=1}^{M_i} \mathcal{K}'_{ij}(\delta) \exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij}) \right\}$$

e

$$U_{\beta_r}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial \beta_r} = \sum_{i=1}^{n} \left\{ d_i \sum_{j=1}^{M_i-1} x_{rij} - \sum_{j=1}^{M_i} x_{rij} \mathcal{K}_{ij}(\delta) \exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij}) \right\},$$

em que $\mathcal{K}'_{ij}(\delta) = (\tilde{y}_{ij} + t_{i,j-1})^\delta \log(\tilde{y}_{ij} + t_{i,j-1}) - t_{i,j-1}^\delta \log(t_{i,j-1})$ e $r = 0, 1, \cdots, p$ com $x_{0ij} = 1$, para todo i e j.

Portanto, $\mathbf{U}(\boldsymbol{\theta})$ é um vetor $\kappa \times 1$, em que κ corresponde ao número de parâmetros do modelo ($\kappa = p + 2$) :

$$\mathbf{U}(\boldsymbol{\theta}) = (U_\delta(\boldsymbol{\theta})\ U_{\beta_0}(\boldsymbol{\theta})\ \cdots\ U_{\beta_p}(\boldsymbol{\theta}))^\top.$$

Os estimadores de máxima verossimilhança (EMVs) de $\boldsymbol{\theta} = (\delta, \beta_0, \cdots, \beta_p)$ podem ser obtidos pela maximização direta da função de log-verossimilhança (3.12), utilizando por exemplo um procedimento de otimização BFGS (Press *et al.*, 2007), ou através da solução das equações dadas por $U_\delta(\boldsymbol{\theta}) = 0$ e $U_{\beta_r}(\boldsymbol{\theta}) = 0$, para $r = 0, 1, \cdots, p$. A solução destas equações não possui forma

fechada, tornando-se necessário também o uso de métodos iterativos para obtenção das estimativas.

Inferências sobre os parâmetros do modelo podem ser baseadas, a princípio, nos EMVs e seus erros padrão estimados. Sob condições de regularidade (Borgan, 1984), podemos aproximar a distribuição do EMV de $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, pela distribuição Normal multivariada com vetor de médias $\boldsymbol{\theta}$ e matriz de covariâncias $\sum(\hat{\boldsymbol{\theta}})$. A matriz de covariâncias é estimada pelo inverso da matriz de informação esperada, a qual pode ser aproximada pela matriz de informação observada, $\mathbf{J}(\boldsymbol{\theta})$. Dessa forma, a matriz $\mathbf{J}(\boldsymbol{\theta})$, com dimensão $\kappa \times \kappa$, pode ser escrita como

$$\mathbf{J}(\hat{\boldsymbol{\theta}}) = -\left.\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{3.13}$$

cujos elementos são dados por

$$J_{\delta\delta}(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial \delta^2} = -\frac{M}{\delta^2} - \sum_{i=1}^{n}\sum_{j=1}^{M_i} \mathcal{K}''_{ij}(\delta)\exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij}),$$

$$J_{\delta\beta_r}(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial \delta \partial \beta_r} = -\sum_{i=1}^{n}\sum_{j=1}^{M_i} \mathrm{x}_{rij}\mathcal{K}'_{ij}(\delta)\exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij})$$

e

$$J_{\beta_r\beta_s}(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial \beta_r \partial \beta_s} = -\sum_{i=1}^{n}\sum_{j=1}^{M_i} \mathrm{x}_{rij}\mathrm{x}_{sij}\mathcal{K}_{ij}(\delta)\exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij}),$$

avaliados em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, em que $\mathcal{K}''_{ij}(\delta) = (\tilde{y}_{ij} + t_{i,j-1})^\delta\{\log(\tilde{y}_{ij} + t_{i,j-1})\}^2 - t^\delta_{i,j-1}\{\log(t_{i,j-1})\}^2$ e $r, s = 0, 1, \cdots, p$, com $\mathrm{x}_{0ij} = 1$, para todo i e j.

A matriz $\mathbf{J}(\boldsymbol{\theta})$ é então expressa na forma

$$\mathbf{J}(\hat{\boldsymbol{\theta}}) = -\begin{pmatrix} J_{\delta\delta}(\boldsymbol{\theta}) & J_{\delta\beta_0}(\boldsymbol{\theta}) & \cdots & J_{\delta\beta_p}(\boldsymbol{\theta}) \\ J_{\delta\beta_0}(\boldsymbol{\theta}) & J_{\beta_0\beta_0}(\boldsymbol{\theta}) & \cdots & J_{\beta_0\beta_p}(\boldsymbol{\theta}) \\ \vdots & \vdots & \ddots & \vdots \\ J_{\delta\beta_p}(\boldsymbol{\theta}) & J_{\beta_0\beta_p}(\boldsymbol{\theta}) & \cdots & J_{\beta_p\beta_p}(\boldsymbol{\theta}) \end{pmatrix},$$

avaliada em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

O intervalo de confiança assintótico, com nível de confiança $100 \times (1-\alpha)\%$, para o k-ésimo componente do vetor de parâmetros $\boldsymbol{\theta}$, θ_k, $k = 1, 2, \cdots, \kappa$ pode ser calculado utilizando

$$\hat{\theta}_k \pm \mathcal{Z}_{\alpha/2} \sqrt{J^{-1}_{(k)}(\hat{\boldsymbol{\theta}})}, \tag{3.14}$$

em que $\mathcal{Z}_{\alpha/2}$ é o valor do $(\alpha/2)$-ésimo quantil superior da distribuição Normal padrão e $J^{-1}_{(k)}(\hat{\boldsymbol{\theta}})$ é o k-ésimo elemento da diagonal principal da inversa da matriz $\mathbf{J}(\hat{\boldsymbol{\theta}})$, que corresponde ao estimador da variância do parâmetro de interesse.

Outra forma de construção dos intervalos de confiança pode ser baseada no método *bootstrap* (Monaco *et al.*, 2005) para dados de sobrevivência multivariados. Este método é uma técnica de reamostragem utilizada para aproximar a distribuição teórica de uma variável aleatória por sua distribuição empírica. No processo de reamostragem pode ser considerada uma especificação paramétrica ou uma não-paramétrica, sendo esta última utilizada para obtenção das estimativas intervalares via *bootstrap*. Neste método, os múltiplos tempos de um mesmo indivíduo são selecionados simultaneamente. Os intervalos de confiança *bootstrap* não-paramétrico são obtidos simulando B amostras com reposição de tamanho n dos dados originais, $\mathcal{D}^*_{(1)}, \mathcal{D}^*_{(2)}, \cdots, \mathcal{D}^*_{(B)}$. Para cada reamostra $\mathcal{D}^*_{(b)}$, $b = 1, \cdots, B$, são calculados os estimadores de máxima verossimilhança dos parâmetros. Os intervalos com nível $100 \times (1-\alpha)\%$ de confiança para cada um dos parâmetros são então obtidos calculando-se os quantis $(1-\alpha/2)$ e $(\alpha/2)$ dos respectivos B estimadores de máxima verossimilhança.

3.2.2 Avaliação do ajuste e seleção de modelos

A escolha de um modelo apropriado para representar um conjunto de dados é um tópico extremamente importante na análise estatística. O modelo proposto engloba o processo de Poisson homogêneo como caso particular,

sendo então natural verificar se um modelo mais simples pode ser considerado. Um método para tratar essa questão consiste em testar a adequação do PPH sob o modelo completo. As hipóteses para testar o PPH é dada como

$$H_0 : \delta = 1 \qquad \text{vs} \qquad H_1 : \delta \neq 1. \tag{3.15}$$

Diversas metodologias para analisar a adequabilidade de modelos são apresentadas na literatura. Dentre elas consideramos duas técnicas baseadas em verossimilhança, sendo: a estatística da razão de verossimilhanças (RV) e a estatística do escore.

Seja $\hat{\boldsymbol{\theta}} = \arg\max_{(\delta,\boldsymbol{\beta})} \ell(\boldsymbol{\theta}|\mathcal{D})$ o EMV obtido a partir do ajuste do modelo completo, sob a hipótese irrestrita, e $\hat{\boldsymbol{\theta}}_0 = \arg\max_{(\delta=1,\boldsymbol{\beta})} \ell(\boldsymbol{\theta}|\mathcal{D})$ o correspondente EMV obtido sob a hipótese restrita H_0. A estatística RV é então dada por

$$\mathrm{X}^2 = 2\{\ell(\hat{\boldsymbol{\theta}}|\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}}_0|\mathcal{D})\}, \tag{3.16}$$

em que $\ell(\cdot)$ denota a função de log-verossimilhança.

A estatística do escore, apresentada por Peng & Xu (2012), é expressa como

$$\mathrm{Z}^2 = \left(\frac{\partial \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial \delta}\right)^2 \Bigg/ \left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial \delta^2} - AB^{-1}A^\top\right)\Bigg|_{\hat{\boldsymbol{\theta}}_0}, \tag{3.17}$$

em que $A = \left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\delta\partial\beta_0^\top}, -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\delta\partial\beta_1^\top}, \cdots, -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\delta\partial\beta_p^\top}\right)$ e

$$B = \begin{pmatrix} -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\beta_0\partial\beta_0^\top} & -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\beta_0\partial\beta_1^\top} & \cdots & -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\beta_0\partial\beta_p^\top} \\ -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\beta_1\partial\beta_0^\top} & -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\beta_1\partial\beta_1^\top} & \cdots & -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\beta_1\partial\beta_p^\top} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\beta_p\partial\beta_0^\top} & -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\beta_p\partial\beta_1^\top} & \cdots & -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\beta_p\partial\beta_p^\top} \end{pmatrix}.$$

Para a hipótese H_0 em (3.15), $\delta = 1$ é um ponto interior do espaço paramétrico de δ. Portanto, a distribuição assintótica de ambas as estatísticas, X^2 e Z^2, segue uma distribuição qui-quadrado com 1 grau de liberdade quando H_0 é verdadeira. Valores positivos elevados de X^2 e Z^2 fornecem evidência

favorável à hipótese H_1. Algumas propriedades amostrais das distribuições dos testes são exploradas via estudo de simulação.

Por fim, uma ferramenta de diagnóstico é discutida a fim de avaliar o ajuste do modelo e garantir que as suposições acerca do mesmo sejam plausíveis aos dados disponíveis. Os resíduos de Cox-Snell são úteis para verificar o ajuste global de um modelo final (Cook & Lawless, 2007). Sendo assim, para o caso de diversos processos recorrentes, $i = 1, \cdots, n$, os resíduos de Cox-Snell são definidos como

$$\hat{r}_{ij} = \int_0^{\tilde{y}_{ij}} \hat{\lambda}_i(u|t_{i,j-1}, \mathbf{x}_{ij})du, \tag{3.18}$$

em que $j = 1, \cdots, M_i$ e $\hat{\lambda}_i(\cdot)$ é a função de taxa obtida do modelo ajustado.

Os resíduos de Cox-Snell para o modelo proposto são expressos na forma

$$\hat{r}_{ij} = \exp(\mathbf{x}_{ij}^\top \hat{\boldsymbol{\beta}}) \left\{ (\tilde{y}_{ij} + t_{i,j-1})^{\hat{\delta}} - t_{i,j-1}^{\hat{\delta}} \right\}, \tag{3.19}$$

em que $\hat{\delta}$ e $\hat{\boldsymbol{\beta}}$ são os EMVs de δ e $\boldsymbol{\beta}$.

Os resíduos r_{ij} são resíduos parciais, ou seja, calcula-se um resíduo para cada recorrência de cada indivíduo. Dessa forma, como os indivíduos têm múltiplos registros, para avaliar o ajuste global do modelo é necessário utilizar um resíduo acumulado. Assim, propomos que estes resíduos acumulados sejam calculados em cada recorrência, a partir dos resíduos r_{ij}, em que na última recorrência de cada indivíduo é registrado um resíduo acumulado total. O gráfico dos resíduos de Cox-Snell é uma ferramenta de diagnóstico visual importante e que permite uma verificação mais direta do modelo. Se o modelo estiver correto, r_{ij} deve se comportar como uma amostra censurada de uma distribuição Exponencial com parâmetro um. Assim, os pontos no gráfico da função de taxa acumulada estimada dos resíduos versus os resíduos de Cox-Snell devem seguir, aproximadamente, uma reta que passa pela origem com inclinação de 45° (Collett, 2003).

Ainda, como uma ligeira modificação dos resíduos de Cox-Snell, pode-se considerar os resíduos de martingale (Cook & Lawless, 2007). Estes resíduos

são definidos, para $i = 1, \cdots, n$, por

$$\hat{m}_{ij} = dN_i(t) + \hat{r}_{ij}, \tag{3.20}$$

em que $dN_i(t)$ é o número de eventos no intervalo $(t_{i,j-1}, t_{i,j-1} + y_{ij}]$ e $\hat{r}_{ij}$ são os resíduos de Cox-Snell dados pela expressão (3.18). Para indicação de um bom ajuste, espera-se que os resíduos de martingale estejam distribuídos aleatoriamente em torno de zero, existindo duas nuvens de pontos, uma representando os indivíduos que falharam e a outra representando indivíduos com tempos censurados.

3.3 Estudo de simulação

Nesta seção consideramos um estudo de simulação para avaliar o desempenho do método de máxima verossimilhança no processo de estimação dos parâmetros e suas propriedades assintóticas, com os seguintes objetivos de estudo: (i) investigar o efeito do tamanho da amostra (n); (ii) investigar o efeito da utilização de uma função de taxa basal crescente ou decrescente; e (iii) investigar a qualidade dos EMVs e seus desvios padrão estimados através da matriz de informação observada. O estudo de simulação também tem como objetivo investigar a distribuição assintótica dos testes da razão de verossimilhanças e do escore sob a hipótese nula, bem como o poder para detectar a hipótese alternativa. O comportamento dos resíduos de Cox-Snell para um conjunto de dados artificial também é analisado.

3.3.1 Dados simulados

A geração da amostra envolve n indivíduos (ou unidades), sendo que cada indivíduo i, $i = 1, \cdots, n$, pode experimentar M_i recorrências do evento. Por simplicidade, duas covariáveis fixas são consideradas, x_{1i} e x_{2i}. Os valores da covariável x_{1i} provêm de uma distribuição de Bernoulli com parâmetro $0,5$,

enquanto os valores da covariável x_{2i} provêm de uma distribuição Normal com média 0 e desvio padrão 2. O tempo de censura C_i para cada indivíduo é gerado a partir de duas distribuições, sendo uma distribuição Uniforme no intervalo $(0, 6)$ e a outra uma distribuição Exponencial com média 3. Os dados de sobrevivência recorrentes são, então, simulados como segue. O tempo final de estudo é fixado em $\tau = 5,9$ anos e, para cada indivíduo i, o tempo de censura C_i é gerado como uma realização independente de uma das distribuições especificadas acima (Uniforme ou Exponencial). Assim, para um indivíduo i temos $\tau_i = \min(\tau, C_i)$ a duração do tempo de acompanhamento, com $[0, \tau_i]$ o período de tempo em que o mesmo é observado. Com isso, os tempos entre eventos y_{ij} são gerados a partir do modelo (3.11), com função de sobrevivência condicional dada por $S(y_{ij}|t_{i,j-1}) = \exp\left\{\left[t_{i,j-1}^{\delta} - (y_{ij} + t_{i,j-1})^{\delta}\right] \mathrm{e}^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}}\right\}$, até que $\sum_{j=1}^{M_i} y_{ij} \geq \tau_i$.

Os valores dos coeficientes de regressão são especificados como $\beta_0 = -0,9$, $\beta_1 = 1,4$ e $\beta_2 = -0,5$. São considerados dois valores para o parâmetro δ, com $\delta \in \{0,7;\ 1,2\}$, os quais correspondem a formas decrescente e crescente, respectivamente, da função de taxa basal. Dessa forma, duas configurações são investigadas: (a) Grupo I ($\delta = 0,7$, $\beta_0 = -0,9$, $\beta_1 = 1,4$, $\beta_2 = -0,5$) e (b) Grupo II ($\delta = 1,2$, $\beta_0 = -0,9$, $\beta_1 = 1,4$, $\beta_2 = -0,5$). O vetor de parâmetros é denotado por $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \delta)$. Os valores atribuídos a cada parâmetro, bem como as distribuições consideradas para o tempo de censura foram escolhidos de tal forma a produzir uma quantidade razoável de observações para cada unidade na amostra gerada, possibilitando avaliar adequadamente o desempenho do método de máxima verossimilhança. A implementação computacional foi desenvolvida no *software* R (R Core Team, 2014). As estimativas de máxima verossimilhança são obtidas pela maximização direta da função de log-verossimilhança utilizando um procedimento BFGS, mais especificamente a rotina `optim` do R.

A seguir, um estudo de simulação com apenas uma amostra é considerado

para ilustrar o processo de estimação e o comportamento dos resíduos de Cox-Snell para o Modelo I proposto.

Estimação

Uma amostra de tamanho $n = 200$ é gerada baseada no procedimento discutido no início desta seção. As estimativas de máxima verossimilhança e os respectivos intervalos de confiança de 95% assintótico (utilizando a matriz de informação observada) são obtidos de acordo com o procedimento inferencial discutido na Seção 3.2. Além disso, um procedimento *bootstrap* não-paramétrico é realizado, considerando $B = 499$ réplicas, obtendo assim os intervalos *bootstrap* com 95% de confiança. Estes resultados são apresentados na Tabela 3.1. Os intervalos de confiança assintóticos são denotados por IC_a enquanto os intervalos *bootstrap* são denotados por IC_b.

Tabela 3.1: Estimativas de máxima verossimilhança e intervalos de confiança de 95% para os parâmetros do modelo a partir dos dados simulados.

		Distribuição do tempo de censura					
		Uniforme			Exponencial		
	Parâmetro	EMV	IC_a (95%)	IC_b (95%)	EMV	IC_a (95%)	IC_b (95%)
(a) Grupo I							
	β_0	-0,907	(-1,106; -0,709)	(-1,114; -0,681)	-0,957	(-1,184; -0,730)	(-1,270; -0,771)
	β_1	1,428	(1,234; 1,621)	(1,235; 1,578)	1,470	(1,248; 1,693)	(1,243; 1,704)
	β_2	-0,501	(-0,538; -0,464)	(-0,536; -0,462)	-0,473	(-0,509; -0,436)	(-0,511; -0,444)
	δ	0,669	(0,623; 0,714)	(0,630; 0,722)	0,712	(0,663; 0,761)	(0,671; 0,768)
(b) Grupo II							
	β_0	-0,849	(-1,010; -0,687)	(-1,016; -0,688)	-0,962	(-1,145; -0,779)	(-1,192; -0,776)
	β_1	1,386	(1,247; 1,525)	(1,246; 1,518)	1,375	(1,209; 1,541)	(1,234; 1,575)
	β_2	-0,503	(-0,534; -0,472)	(-0,535; -0,477)	-0,495	(-0,541; -0,450)	(-0,562; -0,442)
	δ	1,184	(1,126; 1,242)	(1,138; 1,243)	1,250	(1,180; 1,320)	(1,173; 1,316)

É possível observar, a partir dos resultados dispostos na Tabela 3.1, que os intervalos de confiança assintótico e *bootstrap* não apresentam grandes diferenças e, em todas as situações, estes intervalos contêm o verdadeiro valor dos parâmetros. No geral, as estimativas de todos os parâmetros envolvidos no modelo foram satisfatórias quando comparadas aos seus verdadeiros valores.

Resíduos de Cox-Snell

A qualidade do ajuste do modelo é avaliada, com base nos dados simulados, por meio dos resíduos de Cox-Snell discutidos na Seção 3.2.2. A Figura 3.1 apresenta os gráficos com os ajustes globais de Cox-Snell para as situações investigadas. Um bom ajuste é observado quando os pontos gerados seguem próximos à diagonal. Os resíduos de Cox-Snell dispostos na Figura 3.1 estão bem próximos da linearidade para todos os casos, não apresentando nenhum desvio significativo, confirmando assim que não há evidência de falta de ajuste do modelo proposto.

Ainda, com o objetivo de verificar o comportamento dos resíduos de um modelo mal especificado, os resíduos de Cox-Snell foram calculados considerando-se o ajuste do PPH para os dados simulados do modelo completo. A Figura 3.2 apresenta os gráficos com os ajustes globais de Cox-Snell para o modelo PPH.

É possível observar, a partir da Figura 3.2, que os resíduos de Cox-Snell para o modelo PPH apresentam desvios da linearidade e, quando comparados aos resíduos para o modelo completo evidenciam a inadequação do modelo PPH. Portanto, os resultados das simulações mostram que o procedimento de diagnóstico considerado é válido para avaliar a adequação do modelo, além de mostrar a eficácia da forma de geração dos dados bem como a eficácia do método de estimação dos parâmetros na presença de censura.

3.3.2 Propriedades frequentistas dos estimadores de máxima verossimilhança e performance dos testes de hipóteses

Para examinar as propriedades frequentistas dos estimadores de máxima verossimilhança, construímos os intervalos de confiança assintótico e *bootstrap*

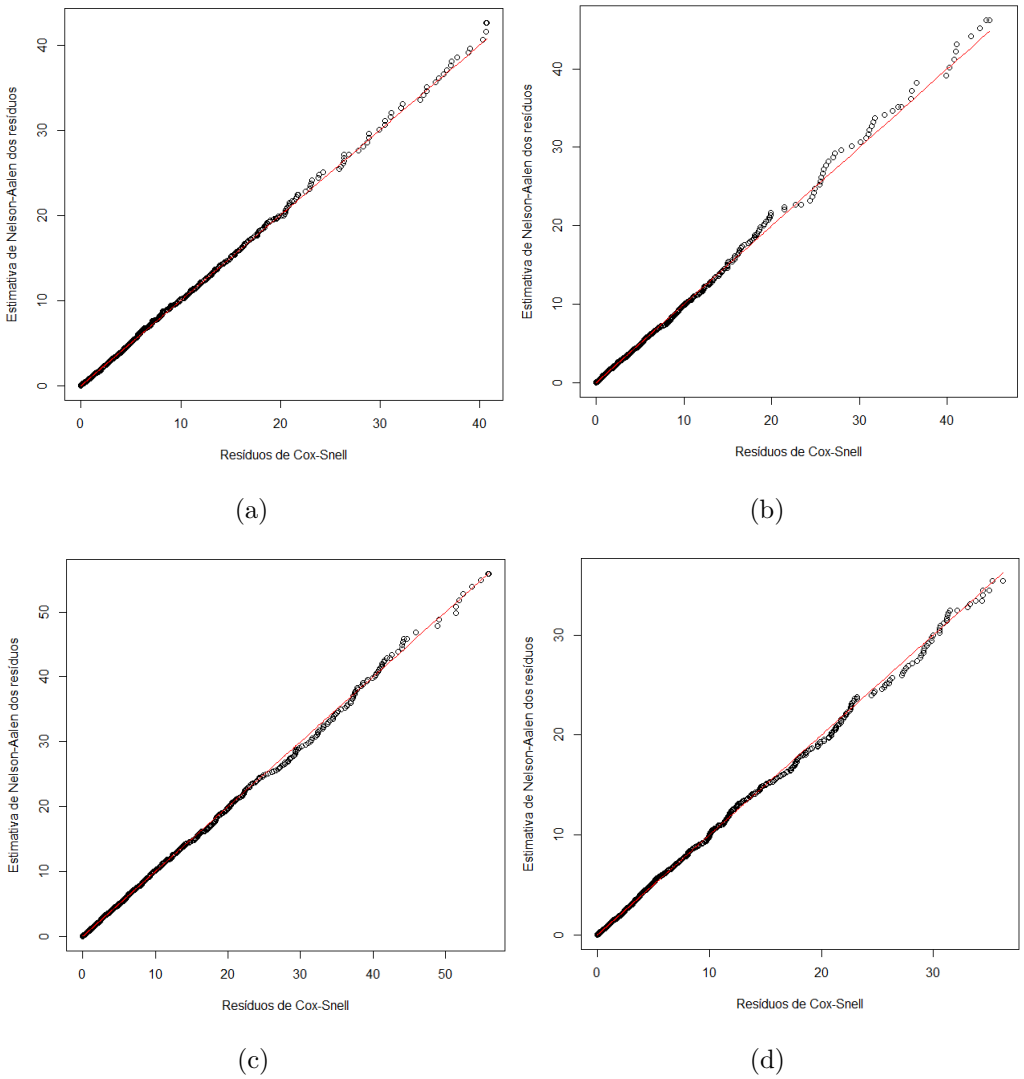


Figura 3.1: Resíduos de Cox-Snell ajustados para o modelo completo. Dados simulados do modelo completo considerando (a) Grupo I e tempo de censura Uniforme, (b) Grupo I e tempo de censura Exponencial, (c) Grupo II e tempo de censura Uniforme e (d) Grupo II e tempo de censura Exponencial.

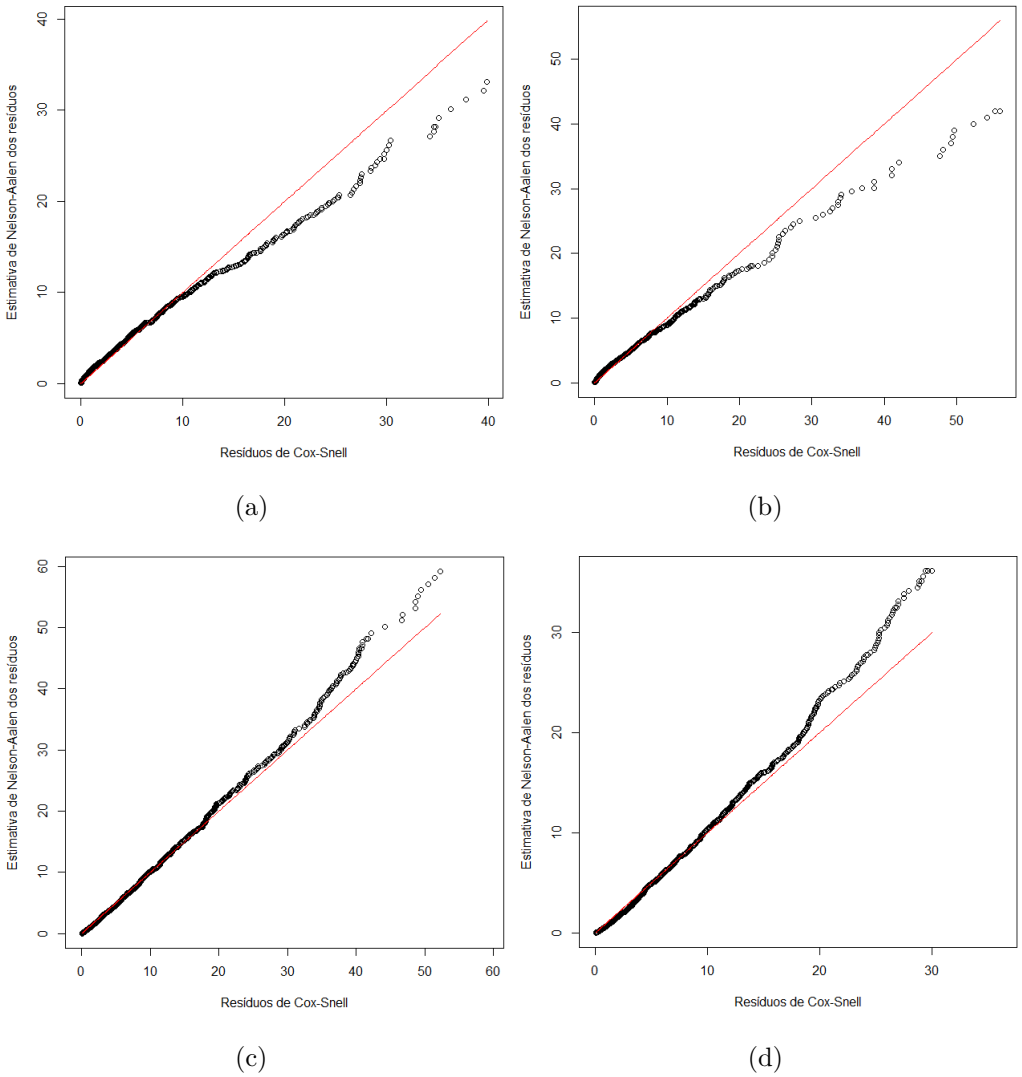


Figura 3.2: Resíduos de Cox-Snell ajustados para o modelo PPH. Dados simulados do modelo completo considerando (a) Grupo I e tempo de censura Uniforme, (b) Grupo I e tempo de censura Exponencial, (c) Grupo II e tempo de censura Uniforme e (d) Grupo II e tempo de censura Exponencial.

para os parâmetros e calculamos suas probabilidades de cobertura (PC). As estimativas das probabilidades de cobertura dos intervalos de confiança foram construídas para o nível de confiança fixado em 95%. A determinação das estimativas das probabilidades de cobertura foram obtidas calculando-se a proporção de intervalos que continham o verdadeiro valor dos parâmetros fixados na geração dos dados, baseada em processo de simulação similar ao descrito na Seção 3.3.1. O cálculo da proporção está baseado na simulação de 1.000 amostras de tamanhos $n = 100,\ 200$ e 500. Para os intervalos de confiança *bootstrap* foram consideradas $B = 499$ reamostragens para cada amostra simulada. Além disso, para avaliar a eficiência do estimador de cada parâmetro, estimativas de Monte Carlo do erro quadrático médio (EQM), do desvio padrão do estimador (DP) e do vício foram calculadas. Nesta avaliação, consideramos a média dos vícios relativos ($\mathcal{B}$) e a razão entre a raiz quadrada do EQM e DP. Para estimadores assintoticamente não viciados é esperado que esta razão se aproxime de um à medida que aumentamos o tamanho do conjunto de dados. As estimativas de Monte Carlo foram obtidas com as seguintes equações:

$$\mathrm{EQM}(\hat{\theta}_k) = \frac{1}{Q}\sum_{q=1}^{Q}(\hat{\theta}_{kq} - \theta_k)^2, \qquad \mathrm{DP}(\hat{\theta}_k) = \left(\frac{1}{Q-1}\sum_{q=1}^{Q}(\hat{\theta}_{kq} - \bar{\theta}_k)^2\right)^{1/2}$$

e

$$\mathcal{B}(\hat{\theta}_k) = \frac{1}{Q}\sum_{q=1}^{Q}\frac{(\hat{\theta}_{kq} - \theta_k)}{\theta_k},$$

em que $\bar{\theta}_k = \frac{1}{Q}\sum_{q=1}^{Q}\hat{\theta}_{kq}$, θ_k é o k-ésimo componente do vetor de parâmetros $\boldsymbol{\theta}$, $\hat{\theta}_k$ o estimador de máxima verossimilhança de θ_k e Q o número de amostras geradas. Os resultados dessa simulação estão resumidos nas Tabelas 3.2 - 3.4. Os histogramas dos parâmetros estimados são apresentados no Apêndice A.

A Tabela 3.2 apresenta as probabilidades de cobertura dos intervalos assintótico (PC_a) e *bootstrap* (PC_b), que em geral não apresentam grandes diferenças. As estimativas das probabilidades de cobertura empíricas estão

próximas do nível de cobertura nominal de 95%, particularmente para os intervalos de confiança assintótico, variando entre $93,3\%$ e $96,4\%$. No caso dos intervalos de confiança *bootstrap* as estimativas das probabilidades de cobertura estão entre 92% e $96,5\%$.

Os valores médios das estimativas pontuais de máxima verossimilhança de cada parâmetro bem como o desvio padrão empírico das estimativas (DP) e a média dos desvios padrão estimados ($\widehat{\text{DP}}$), usando a inversa da matriz de informação observada, são apresentados na Tabela 3.3. Esta tabela também apresenta o número médio de eventos observado por unidade baseado nas 1.000 simulações ($\hat{\mu}_E$). A análise dos resultados da Tabela 3.3 permite concluir que o método de máxima apresenta um bom desempenho na obtenção das estimativas pontuais dos parâmetros, uma vez que estas estimativas para todos os parâmetros foram bastante satisfatórias quando comparadas aos seus verdadeiros valores. As estimativas do desvio padrão obtidas da matriz de informação observada são satisfatórias. As médias dos desvios padrão estimados estão bem próximas dos desvios padrão calculados empiricamente, mostrando assim a precisão e relevância das estimativas. Ainda, como esperado, os desvios padrão das estimativas decrescem quando o número de observações na amostra aumenta.

A Tabela 3.4 apresenta as estimativas de Monte Carlo de algumas medidas de eficiência dos estimadores de máxima verossimilhança e o vício médio dos estimadores de cada parâmetro. É possível observar que a média das razões entre a raiz quadrada do EQM e o DP de cada estimador estão próximas de um, indicando que os estimadores dos parâmetros são consistentes assintoticamente. Além disso, para todos os parâmetros, os vícios dos estimadores são pequenos e tornam-se ainda mais próximos de zero quando o tamanho da amostra é aumentado. Com isso, pode-se concluir que os estimadores de δ, β_0 β_1 e β_2 são empiricamente não-viciados. Ainda, a utilização de uma função de taxa basal crescente ou decrescente, bem como a distribuição utilizada para o

Tabela 3.2: Estimativas das probabilidades de cobertura dos intervalos assintótico (PC_a) e *bootstrap* (PC_b) para os parâmetros do modelo.

		Distribuição do tempo de censura			
		Uniforme		Exponencial	
n	Parâmetro	PC_a	PC_b	PC_a	PC_b
(a) Grupo I					
100	β_0	0,940	0,952	0,954	0,920
	β_1	0,943	0,948	0,951	0,940
	β_2	0,942	0,940	0,957	0,926
	δ	0,933	0,960	0,951	0,944
200	β_0	0,951	0,952	0,957	0,944
	β_1	0,958	0,936	0,956	0,960
	β_2	0,955	0,956	0,958	0,936
	δ	0,958	0,944	0,949	0,940
500	β_0	0,952	0,947	0,957	0,949
	β_1	0,958	0,944	0,955	0,965
	β_2	0,953	0,950	0,954	0,946
	δ	0,943	0,940	0,943	0,939
(b) Grupo II					
100	β_0	0,952	0,940	0,949	0,944
	β_1	0,953	0,940	0,953	0,964
	β_2	0,949	0,960	0,956	0,952
	δ	0,941	0,932	0,953	0,936
200	β_0	0,956	0,924	0,952	0,936
	β_1	0,957	0,928	0,958	0,932
	β_2	0,948	0,933	0,964	0,952
	δ	0,952	0,936	0,958	0,960
500	β_0	0,959	0,944	0,945	0,940
	β_1	0,954	0,946	0,949	0,939
	β_2	0,941	0,944	0,948	0,940
	δ	0,950	0,945	0,957	0,945

tempo de censura não interferem nos resultados.

Para avaliar a viabilidade das aproximações das distribuições dos testes,

Tabela 3.3: Médias e respectivos desvios padrão das estimativas de máxima verossimilhança dos parâmetros do modelo baseado nas 1.000 simulações.

		Distribuição do tempo de censura					
		Uniforme			Exponencial		
n	Parâmetro	Média	DP	Média($\widehat{\text{DP}}$)	Média	DP	Média($\widehat{\text{DP}}$)
(a) Grupo I							
100	β_0	-0,908	0,151	0,144	-0,909	0,154	0,153
	β_1	1,406	0,146	0,141	1,404	0,153	0,150
	β_2	-0,498	0,032	0,030	-0,500	0,033	0,033
	δ	0,701	0,037	0,035	0,700	0,037	0,036
	$\hat{\mu}_E$	3,489	–	–	3,084	–	–
200	β_0	-0,905	0,098	0,100	-0,907	0,105	0,106
	β_1	1,401	0,097	0,097	1,409	0,101	0,104
	β_2	-0,501	0,020	0,020	-0,499	0,022	0,022
	δ	0,700	0,024	0,025	0,700	0,025	0,025
	$\hat{\mu}_E$	3,508	–	–	3,050	–	–
500	β_0	-0,900	0,063	0,062	-0,904	0,064	0,066
	β_1	1,398	0,060	0,061	1,404	0,064	0,065
	β_2	-0,500	0,012	0,012	-0,500	0,013	0,013
	δ	0,701	0,016	0,016	0,699	0,016	0,016
	$\hat{\mu}_E$	3,494	–	–	3,054	–	–
(b) Grupo II							
100	β_0	-0,908	0,115	0,115	-0,914	0,124	0,124
	β_1	1,405	0,104	0,102	1,404	0,112	0,112
	β_2	-0,501	0,023	0,022	-0,501	0,025	0,025
	δ	1,199	0,043	0,042	1,202	0,044	0,043
	$\hat{\mu}_E$	6,614	–	–	5,611	–	–
200	β_0	-0,901	0,077	0,080	-0,905	0,085	0,085
	β_1	1,399	0,069	0,071	1,403	0,076	0,077
	β_2	-0,500	0,015	0,015	-0,501	0,016	0,016
	δ	1,200	0,029	0,029	1,200	0,028	0,030
	$\hat{\mu}_E$	6,617	–	–	5,675	–	–
500	β_0	-0,900	0,048	0,050	-0,901	0,053	0,053
	β_1	1,399	0,044	0,044	1,402	0,047	0,048
	β_2	-0,501	0,009	0,009	-0,500	0,010	0,010
	δ	1,199	0,018	0,018	1,199	0,018	0,019
	$\hat{\mu}_E$	6,634	–	–	5,693	–	–

Tabela 3.4: Medidas de eficiência do estimador de cada parâmetro.

		Distribuição do tempo de censura			
		Uniforme		Exponencial	
n	Parâmetro	$\sqrt{EQM}/DP$	$\mathcal{B}$	$\sqrt{EQM}/DP$	$\mathcal{B}$
(a) Grupo I					
100	β_0	$1,001$	$0,0086$	$1,001$	$0,0102$
	β_1	$1,000$	$0,0046$	$1,000$	$0,0028$
	β_2	$1,002$	$-0,0041$	$1,000$	$0,0001$
	δ	$1,000$	$0,0008$	$1,000$	$0,0003$
200	β_0	$1,001$	$0,0056$	$1,002$	$0,0076$
	β_1	$1,000$	$0,0007$	$1,003$	$0,0062$
	β_2	$1,000$	$0,0013$	$1,001$	$-0,0021$
	δ	$1,000$	$0,0005$	$1,000$	$-0,00005$
500	β_0	$1,000$	$-0,0004$	$1,001$	$0,0041$
	β_1	$1,000$	$-0,0014$	$1,001$	$0,0025$
	β_2	$1,000$	$-0,0009$	$1,000$	$-0,00009$
	δ	$1,001$	$0,0013$	$1,000$	$-0,0008$
(b) Grupo II					
100	β_0	$1,002$	$0,0091$	$1,006$	$0,0152$
	β_1	$1,001$	$0,0039$	$1,000$	$0,0032$
	β_2	$1,001$	$0,0024$	$1,001$	$0,0026$
	δ	$1,000$	$-0,0005$	$1,001$	$0,0019$
200	β_0	$1,000$	$0,0015$	$1,002$	$0,0060$
	β_1	$1,000$	$-0,0004$	$1,000$	$0,0022$
	β_2	$1,000$	$0,0010$	$1,000$	$0,0011$
	δ	$1,000$	$0,0003$	$1,000$	$0,0001$
500	β_0	$1,000$	$-0,0002$	$1,000$	$0,0008$
	β_1	$1,000$	$-0,0008$	$1,000$	$0,0013$
	β_2	$1,002$	$0,0012$	$1,000$	$-0,0004$
	δ	$1,000$	$-0,0004$	$1,000$	$-0,0004$

construímos 1.000 amostras de tamanhos 100, 200 e 500 sob a hipótese nula em (3.15) e comparamos as estatísticas RV e do escore com suas respectivas distribuições assintóticas. Analisamos também o poder dos testes para detectar a hipótese alternativa em (3.15). A performance das estatísticas dos testes em (3.16) e (3.17) foi testada considerando-se um nível de significância nominal de 5% na comparação do modelo proposto com seu caso particular (PPH). Os resultados estão organizados na Tabela 3.5, a qual apresenta as proporções empíricas do erro do tipo I e o poder dos testes da razão de verossimilhanças X^2 e do escore Z^2 ao nível de significância nominal de 5%. Pode-se observar que, para ambos os testes, a taxa de rejeição da hipótese nula atingiu o nível de significância esperado teoricamente para os diferentes tamanhos amostrais ($n = 100,\ 200$ e 500). Um maior poder para os testes é observado com o aumento do tamanho da amostra. Quando o tamanho da amostra é 200, o poder de ambos os testes é superior a 90%. Um poder ainda maior é alcançado quando o tamanho da amostra é de 500. Além disso, os resultados apresentados na Tabela 3.5 permitem concluir, como esperado, que a distribuição do tempo de censura não compromete a performance dos testes.

Tabela 3.5: Proporções empíricas do erro do tipo I e poder dos testes da razão de verossimilhanças e escore a um nível de significância de 5%.

		Distribuição do tempo de censura			
		Uniforme		Exponencial	
n	Teste	Erro tipo I	Poder	Erro tipo I	Poder
100	RV	0,054	0,718	0,046	0,676
	Escore	0,056	0,699	0,045	0,663
200	RV	0,050	0,954	0,050	0,934
	Escore	0,051	0,949	0,050	0,933
500	RV	0,046	0,999	0,053	0,999
	Escore	0,046	0,999	0,052	0,999

3.4 Análise de dados reais

Nesta seção demostramos a aplicação do modelo proposto e o procedimento de estimação utilizando um conjunto de dados reais da literatura referente a sucessivas reinternações de pacientes diagnosticados com câncer colorretal. A escolha deste particular conjunto de dados se justifica pelo fato do mesmo permitir uma comparação entre os resultados obtidos com a modelagem proposta neste trabalho e aqueles obtidos anteriormente na literatura a partir de modelagens similares.

O conjunto de dados referente a reinternações hospitalares entre pacientes diagnosticados com câncer colorretal em um estudo de coorte é apresentado em Gonzalez *et al.* (2005). Os pacientes diagnosticados com câncer colorretal entre Janeiro de 1996 e Dezembro de 1998 foram acompanhados ativamente até 2002. Os dados fornecem os tempos entre as sucessivas reinternações (em dias) após a cirugia para remoção do tumor. Um total de 861 reinternações devido à recorrência do câncer colorretal foram registradas entre os 403 pacientes incluídos no estudo. Entre os pacientes, 200 indivíduos $(49,6\%)$ não apresentaram recorrências até o final do estudo. As covariáveis que compõem o conjunto de dados são: quimioterapia (x_1, em que 1: recebeu quimioterapia e 0: caso contrário); sexo (x_2, em que 1: feminino e 0: masculino); estágio do tumor, de acordo com a classificação de Dukes, (x_3, em que 1: estágio A-B, 2: estágio C e 3: estágio D); e índice de comorbidade de Charlson (x_4, em que 0: índice 0, 1: índice $1-2$ e 3: índice ≥ 3). As covariáveis x_1, x_2 e x_3 são consideradas fixas, enquanto a covariável x_4 é modelada como uma covariável externa dependente do tempo.

Segundo Gonzalez *et al.* (2005), a recorrência é o principal risco após a cirurgia e geralmente a última causa de morte. Sendo assim, é de grande importância predizer a probabilidade de recorrência bem como analisar e descrever a frequência de recorrências. O modelo com função de taxa dada por

(3.11) foi então aplicado aos dados de reinternação hospitalar, abordando como covariáveis as quatro informações descritas anteriormente, sendo que para cada indivíduo o vetor de covariáveis associado é dado por $\mathbf{x}_{ij} = (x_{1i}, x_{2i}, x^*_{3i}, x^*_{4ij})^\top$, com $x^*_{3i} = (I(x_{3i} = 2), I(x_{3i} = 3))$ e $x^*_{4ij} = (I(x_{4ij} = 1), I(x_{4ij} = 3))$. Os resultados da análise considerando o modelo proposto estão condensados na Tabela 3.6, a qual apresenta as estimativas de máxima verossimilhança dos parâmetros, bem como seus respectivos desvios padrão (DP), intervalos de confiança de 95% assintótico ($IC_{95\%}$) e valor-p. Os intervalos *bootstrap* são omitidos uma vez que não apresentam grandes diferenças em relação aos intervalos assintóticos. O valor-p associado às covariáveis são calculados considerando-se o teste de Wald. O símbolo $*$ indica que o valor-p é não significativo ao nível de 5%.

Tabela 3.6: Estimativas dos parâmetros do modelo para os dados de reinternação hospitalar.

Descrição	Parâmetro	Estimativa	DP	$IC_{95\%}$	valor-p
Intercepto	β_0	-5,343	0,264	(-5,861; -4,825)	$< 0,001$
Quimio	β_1	-0,197	0,104	(-0,401; 0,007)	$0,100^*$
Sexo (Fem.)	β_2	-0,529	0,101	(-0,726; -0,332)	0,001
Estágio Tumor					
A-B (ref)					
C	β_{3_1}	0,350	0,122	(0,110; 0,590)	0,024
D	β_{3_2}	1,447	0,137	(1,179; 1,715)	$< 0,001$
Índice Charlson					
0 (ref)					
$1-2$	β_{4_1}	0,459	0,205	(0,057; 0,860)	$0,060^*$
≥ 3	β_{4_2}	0,441	0,113	(0,220; 0,663)	0,006
Parâmetro					
Weibull	δ	0,761	0,032	(0,698; 0,825)	$< 0,001$

A partir dos resultados apresentados na Tabela 3.6, é possível observar que o tratamento com quimioterapia e a gravidade do tumor dada pelo índice

de Charlson $1-2$ não têm efeito significativo sobre os tempos de recorrência dos pacientes (valor-p $= 0,100$ e valor-p $= 0,060$, respectivamente). Isso mostra que, embora estas covariáveis tenham sido consideradas como fatores que podem influenciar a recorrência de câncer colorretal, há pouca evidência estatística para apoiar essa suposição. Com isso, os principais fatores de risco para a reinternação devido à recorrência de câncer colorretal são sexo, estágio do tumor e índice de Charlson ≥ 3. Os resultados do ajuste que considera apenas as covariáveis significativas (ao nível de 5%) são apresentados na Tabela 3.7.

Tabela 3.7: Estimativas dos parâmetros do modelo para os dados de reinternação hospitalar considerando apenas as covariáveis significativas.

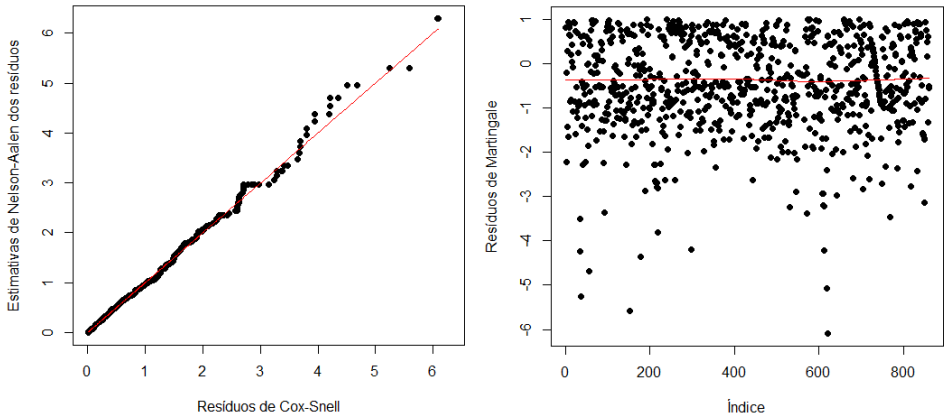
Descrição	Parâmetro	Estimativa	DP	$IC_{95\%}$	valor-p
Intercepto	β_0	-5,534	0,249	(-6,022; -5,046)	$< 0,001$
Sexo (Fem.)	β_2	-0,534	0,101	(-0,731; -0,336)	0,003
Estágio Tumor					
A-B (ref)					
C	β_{3_1}	0,404	0,115	(0,179; 0,630)	0,017
D	β_{3_2}	1,453	0,135	(1,190; 1,717)	$< 0,001$
Índice Charlson					
0 (ref)					
≥ 3	β_{4_2}	0,472	0,112	(0,253; 0,691)	0,008
Parâmetro					
Weibull	δ	0,775	0,033	(0,711; 0,839)	$< 0,001$

Com os resultados da Tabela 3.7 podemos concluir que há evidência de uma diferença significativa entre o sexo feminino e masculino (valor-p $= 0,003$) com relação ao risco de recorrência. O valor negativo de $\hat{\beta}_2$ indica que os tempos entre reinternações para pacientes do sexo feminino são significativamente maiores do que para os pacientes do sexo masculino. Pode-se observar que um paciente do sexo masculino tem um risco 70% maior de apresentar recorrências do que um paciente do sexo feminino, com

risco relativo igual a $1,71$ ($IC_{95\%} = (1,40;\ 2,08)$). Estágios do tumor mais avançados e um alto índice de comorbidade de Charlson estão associados a tempos entre reinternações menores. Estas covariáveis contribuem para o aumento do risco de reinternações devido à recorrência de câncer colorretal, com riscos relativos iguais a $1,50$ ($IC_{95\%} = (1,20;\ 1,88)$) e $4,28$ ($IC_{95\%} = (3,28;\ 5,57)$) para os estágios C e D, respectivamente, e risco relativo igual a $1,60$ ($IC_{95\%} = (1,28;\ 2,00)$) para índice de Charlson ≥ 3.

Para testar a possibilidade de um modelo mais simples, isto é, $H_0 : \delta = 1$ versus $H_1 : \delta \neq 1$, são utilizados os testes RV e do escore apresentados na Seção 3.2.2. Ambos os testes apresentam forte evidência a favor do modelo completo, com ambos valores-p $< 0,0001$, indicando que um modelo com a capacidade de capturar a possível correlação entre os tempos de reinternações dos pacientes é preferível neste caso. A qualidade do ajuste global do modelo é avaliada por meio dos gráficos dos resíduos de Cox-Snell e resíduos de martingale. A Figura 3.3 apresenta os gráficos com os resíduos de Cox-Snell e resíduos de martingale a partir do ajuste do modelo proposto para os dados de reinternação hospitalar. Pelo gráfico dos resíduos de martingale (Figura 3.3(b)) é possível notar alguns indivíduos mal ajustados, o que pode ser observado pelos pontos cujos resíduos são mais negativos comparados aos demais. Uma explicação para este fato pode ser dada com base nos indivíduos que não apresentaram recorrências durante o período de acompanhamento. Para estes indivíduos o tempo de sobrevivência é censurado pelo tempo final do estudo, o qual neste caso é bastante longo, produzindo assim uma estimativa grande do resíduo. De maneira geral, os resíduos apresentados na Figura 3.3 não apresentam nenhum desvio significativo, indicando um bom ajuste do modelo proposto aos dados de reinternação hospitalar.

Por fim, comparamos nossa metodologia e resultados com aqueles de estudos anteriores. Para análise do mesmo conjunto de dados, Gonzalez *et al.* (2005) e Rondeau *et al.* (2011) utilizam um modelo de fragilidade, baseado no



(a) (b)

Figura 3.3: Resíduos ajustados para os dados de reinternação hospitalar. (a) resíduos de Cox-Snell e (b) resíduos de martingale.

modelo de riscos proporcionais de Cox, para explicar a heterogeneidade não observada entre os tempos de um mesmo indivíduo. Nosso modelo, por outro lado, considera cada um dos tempos entre eventos condicionado ao tempo da recorrência anterior, captando assim a correlação existente entre os tempos sem maiores problemas. No caso de câncer colorretal, Gonzalez *et al.* (2005) obtiveram, dentre outras características, evidência de diferença significativa principalmente entre o sexo feminino e masculino, enquanto Rondeau *et al.* (2011) obtiveram efeito significativo das características sexo, estágio do tumor e índice de Charlson. Nossa análise, em comparação, fornece conclusões que estão em concordância com as conclusões obtidas pelos autores acima citados.

3.5 Alguns comentários

Neste capítulo propomos um modelo paramétrico que possibilita investigar os tempos entre sucessivas ocorrências de um evento recorrente com uma dependência induzida pelo tempo de censura quando $M_i > 1,\ i = 1, \cdots, n$. Ao

invés de simplesmente assumir um modelo de riscos multiplicativos ou aditivos para os tempos entre eventos, propomos um modelo em que a distribuição condicional do tempo entre sucessivas recorrências é derivada facilmente da função de taxa marginal, que é uma formulação atrativa para dados de eventos recorrentes e possibilita interpretações práticas mais diretas para a identificação de fatores de risco com base nos coeficientes das covariáveis.

Em geral, os processos de eventos recorrentes violam o pressuposto de independência. Qualquer correlação entre os eventos pode acarretar consequências importantes, tais como estimativas ineficientes e viesadas, e consequentemente estimativas de erros padrão incorretas. Devido a este fato, o modelo é formulado considerando cada um dos tempos entre eventos condicionado ao tempo da recorrência anterior. Com isso, os tempos entre eventos são tratados da mesma forma e a relação entre as sucessivas ocorrências deixa de ser um problema. O modelo proposto é, assim, uma alternativa aos modelos já existentes para tempos entre eventos recorrentes, tais como os modelos de fragilidade.

O modelo proposto abrange como caso particular o clássico modelo de Poisson homogêneo, que pode ser testado diretamente. O procedimento inferencial é baseado na abordagem de máxima verossimilhança, o qual possibilitou fácil implementação sem esforços computacionais. A rotina `optim` do *software* R, que é uma rotina de otimização para fins gerais, foi considerada para a maximização direta da função de log-verossimilhança. Este procedimento tem a vantagem de ser executado rapidamente e de fácil uso. Além disso, este procedimento não apresentou problemas numéricos, tais como problemas de convergência, em nenhum dos estudos realizados.

Os resultados do estudo de simulação mostraram a eficácia do método de estimação dos parâmetros, mesmo para uma quantidade pequena de unidades na amostra, independente da distribuição assumida para o tempo de censura. A importância prática do modelo proposto para tempos entre eventos

recorrentes foi demonstrada em um conjunto de dados reais da literatura, em que a metodologia apresentada e os resultados obtidos a partir da mesma foram comparados com aqueles de estudos anteriores. Além disso, os gráficos dos resíduos de Cox-Snell e de martingale foram de grande importância por permitir uma avaliação mais direta do modelo para dados de eventos recorrentes.

Capítulo 4

Modelo para Tempos entre Eventos Recorrentes com Fragilidade Discreta (Modelo II)

Neste capítulo, apresentamos uma modelagem alternativa para dados de eventos recorrentes que estende os modelos de fragilidade permitindo o uso de distribuições como a Bernoulli, Geométrica, Poisson, Weibull Discreta, Binomial Negativa ou outra distribuição discreta para a variável de fragilidade. As distribuições de fragilidade contínuas não permitem a possibilidade de indivíduos com risco zero, e portanto em situações que esta condição se faz presente uma distribuição discreta pode ser mais apropriada. Um exemplo para tal situação é quando a fragilidade surge devido a presença de um número desconhecido de fatores que levam à ocorrência do evento.

O Modelo II proposto neste capítulo permite a possibilidade de risco zero para a recorrência de um determinado evento de interesse, uma vez que a fragilidade assumindo valor zero corresponde a um modelo que contém uma

proporção de indivíduos que nunca falham, ou seja, indivíduos que tornaram-se imunes ou não susceptíveis ao evento em questão. Com isso, além de estender os modelos de fragilidade para dados de eventos recorrentes, essa metodologia também permite estudar a possibilidade de o indivíduo ser não suscetível ao evento de interesse através de outros modelos para a função de sobrevivência da população.

Para inferir sobre os parâmetros do modelo consideramos uma abordagem clássica, em que as estimativas foram obtidas pelo método de máxima verossimilhança. A função de sobrevivência para os indivíduos susceptíveis segue o modelo apresentado no Capítulo 3, assumindo uma função de taxa basal Weibull. Para avaliar as propriedades dos estimadores consideramos um estudo de simulação. A metodologia proposta foi utilizada para analisar um conjunto de dados de reinternações hospitalares entre pacientes diagnosticados com câncer colorretal.

4.1 Formulação do modelo

Suponha que um indivíduo sob estudo pode experimentar consecutivas recorrências de um mesmo tipo de evento nos tempos $0 < T_1 < T_2 < \cdots < T_j < \cdots$, $j = 1, 2, \cdots$, medidos a partir do início do estudo. Os tempos entre eventos são então definidos como $Y_j = T_j - T_{j-1}$, tempo entre o $(j-1)$-ésimo e j-ésimo evento, para $j = 1, 2, \cdots$ e $T_0 = 0$.

Seja Z uma variável aleatória discreta com suporte $\{0, 1, 2, \cdots\}$ usada para modelar a fragilidade. Considerando o Modelo I apresentado no Capítulo 3, a função de taxa do processo de recorrência para um indivíduo até o tempo $y + t$ com fragilidade $Z = z$ e vetor de covariáveis $\mathbf{x}$ é dada por

$$\lambda(y|t, \mathbf{x}, z) = z\lambda(y|t, \mathbf{x}), \tag{4.1}$$

em que $\lambda(y|t, \mathbf{x}) = \lambda_0(y + t)\exp(\boldsymbol{\beta}^\top \mathbf{x})$, como expresso por (3.1), sendo λ_0

uma função basal contínua que descreve o comportamento geral do indivíduo ao longo do tempo e $\boldsymbol{\beta}$ o vetor de coeficientes de regressão associado à $\mathbf{x}$.

O modelo com fragilidade (4.1) também pode ser representado por sua função de sobrevivência, condicionada a $Z = z$, expressa como

$$\begin{aligned} S(y|t,z) &= \exp\left\{-z\int_0^y \lambda(u|t,\mathbf{x})du\right\} \\ &= \exp\{-z\Lambda(t,y)\} = S(y|t)^z, \end{aligned} \tag{4.2}$$

em que $S(y|t) = \Pr(Y_j > y|T_{j-1} = t) = \exp\{-\Lambda(t,y)\}$, como expresso por (3.4), e $\Lambda(t,y) = \int_0^y \lambda(u|t,\mathbf{x})du$ a função de taxa basal acumulada sobre o intervalo $(t, t+y]$.

Assumindo que a distribuição de probabilidade de Z seja especificada por $\Pr\{Z = z\} = p_z$, para $z = 0, 1, 2, \cdots$ com $\sum_{z=0}^{\infty} p_z = 1$, e usando a expressão em (2.10), a função de sobrevivência incondicional, com relação a distribuição de fragilidade discreta, para o modelo (4.1) pode ser escrita como

$$S^*(y|t) = \mathrm{E}\{S(y|t)^Z\} = \sum_{z=0}^{\infty} p_z S(y|t)^z = G_Z(S(y|t)), \tag{4.3}$$

em que $G_Z(\cdot)$ é a função geradora de probabilidade da variável aleatória Z, a qual converge quando $S(y|t) \in [0,1]$.

A função densidade de probabilidade associada à (4.3) é dada por

$$f^*(y|t) = -\frac{\mathrm{d}}{\mathrm{d}y}S^*(y|t) = G_Z'(S(y|t))f(y|t), \tag{4.4}$$

em que $G_Z'(S(y|t)) = \mathrm{d}G_Z(s)/\mathrm{d}s|_{s=S(y|t)}$ e $f(y|t) = -\mathrm{d}S(y|t)/\mathrm{d}y$.

Para dados de sobrevivência na presença de eventos recorrentes, um indivíduo pode apresentar várias recorrências do evento ou nenhuma recorrência. Dentre os indivíduos que não apresentam recorrências, pode haver uma parcela de indivíduos para os quais o evento não se manifestará (indivíduos não susceptíveis ou curados) e, independente do tempo o qual forem acompanhados, não irão mais experimentar o evento, enquanto os indivíduos susceptíveis (não curados) podem ter múltiplas recorrências. O modelo proposto neste

capítulo permite trabalhar com este tipo de situação. A fragilidade igual a zero, ou seja, $Z = 0$, corresponde ao indivíduo não experimentar a recorrência do evento, de modo que o seu tempo até o primeiro evento pode ser infinito. Assim, se Y_1 denota o tempo até o primeiro evento, então Y_1 é infinito se $Z = 0$ com $\Pr\{Y_1 = \infty | Z = 0\} = 1$. Então, o modelo com fragilidade discreta (4.1) permite incorporar uma proporção ($p_0 > 0$) de indivíduos que nunca falham, ou seja, indivíduos não susceptíveis a recorrência do evento.

4.1.1 Casos especiais

O modelo de fragilidade proposto (Modelo II) abrange como casos especiais alguns modelos listados abaixo. Diferentes distribuições discretas podem ser utilizadas para modelar a fragilidade, no entanto neste trabalho, a princípio, são consideradas para a fragilidade Z apenas as distribuições discretas padrão, como a distribuição de Bernoulli, Geométrica e Poisson.

(i) Modelo com fragilidade Bernoulli

Quando Z é uma variável aleatória com distribuição de Bernoulli, com parâmetro $\pi \in (0, 1)$ e $p_z = \pi^z(1 - \pi)^{1-z}$, para $z = 0, 1$, segue de (4.3) que a função de sobrevivência da população é dada por

$$S^*(y|t) = 1 - \pi + \pi S(y|t). \tag{4.5}$$

A correspondente função densidade de probabilidade é dada por

$$f^*(y|t) = \pi f(y|t). \tag{4.6}$$

A proporção de indivíduos não susceptíveis, neste caso, é dada por $p_0 = \Pr\{Z = 0\} = 1 - \pi$, enquanto $1 - p_0 = \Pr\{Z = 1\} = \pi$ denota a proporção de indivíduos susceptíveis (recorrentes).

Note que o modelo (4.5) é o modelo de mistura padrão de Berkson & Gage (1952) para uma estrutura de dados de eventos recorrentes.

(*ii*) Modelo com fragilidade Geométrica

No caso em que Z tem distribuição Geométrica, com parâmetro $\pi \in (0,1)$ e $p_z = \pi^z(1-\pi)$, para $z = 0, 1, 2, \cdots$, a função de sobrevivência da população é dada por

$$S^*(y|t) = (1-\pi)/\{1-\pi S(y|t)\}. \tag{4.7}$$

A correspondente função densidade de probabilidade é dada por

$$f^*(y|t) = (1-\pi)\pi f(y|t)/\{1-\pi S(y|t)\}^2. \tag{4.8}$$

A proporção de indivíduos não susceptíveis, neste caso, é expressa como $p_0 = \Pr\{Z = 0\} = 1-\pi$ e a proporção de indivíduos susceptíveis é dada por $1 - p_0 = \Pr\{Z > 0\} = \pi$.

(*iii*) Modelo com fragilidade Poisson

Por fim, no caso em que Z assume distribuição de Poisson, com parâmetro $\pi > 0$ e $p_z = \mathrm{e}^{-\pi}\pi^z/z!$, para $z = 0, 1, 2, \cdots$, a função de sobrevivência da população é dada por

$$S^*(y|t) = \exp\{-\pi(1-S(y|t))\}. \tag{4.9}$$

A proporção de indivíduos não susceptíveis, neste caso, é dada como $p_0 = \exp(-\pi)$, enquanto $1-p_0 = 1-\exp(-\pi)$ representa a proporção de indivíduos susceptíveis. A correspondente função densidade de probabilidade é dada por

$$f^*(y|t) = \pi f(y|t)\exp\{-\pi(1-S(y|t))\}, \tag{4.10}$$

em que $S(y|t)$ e $f(y|t)$ são, respectivamente, a função de sobrevivência (própria) e a função densidade de probabilidade dos indivíduos recorrentes.

A Tabela 4.1 apresenta as funções de sobrevivência e densidade da população correspondentes aos casos específicos, bem como a probabilidade de ser não suscetível.

Tabela 4.1: Função de sobrevivência da população ($S^*(y|t)$), função densidade ($f^*(y|t)$) e proporção de não suscetíveis (p_0) para os diferentes casos especiais.

Fragilidade	$S^*(y\|t)$	$f^*(y\|t)$	p_0
Bernoulli	$1-\pi+\pi S(y\|t)$	$\pi f(y\|t)$	$1-\pi$
Geométrica	$(1-\pi)/\{1-\pi S(y\|t)\}$	$(1-\pi)\pi f(y\|t)/\{1-\pi S(y\|t)\}^2$	$1-\pi$
Poisson	$\exp\{-\pi(1-S(y\|t))\}$	$\pi f(y\|t)\exp\{-\pi(1-S(y\|t))\}$	$\exp(-\pi)$

A probabilidade de ser não suscetível (ou suscetível) pode variar de indivíduo para indivíduo, uma vez que é razoável assumir que tal probabilidade pode depender de características individuais (covariáveis). Neste sentido, a probabilidade de ser não suscetível pode ser modelada, para os modelos em (i) e (ii), por exemplo, por um função logística, de modo que

$$p_{0_i} = 1 - \pi(\boldsymbol{\omega}_i) = \Pr\{Z_i = 0|\boldsymbol{\omega}_i\} = \frac{\exp(\boldsymbol{b}^\top \boldsymbol{\omega}_i)}{1+\exp(\boldsymbol{b}^\top \boldsymbol{\omega}_i)}, \tag{4.11}$$

em que $\boldsymbol{\omega}_i$ é o vetor de covariáveis e $\boldsymbol{b}$ é o vetor de coeficientes associado à $\boldsymbol{\omega}_i$. O complementar, $1 - p_{0_i} = \pi(\boldsymbol{\omega}_i)$, que representa a probabilidade de um indivíduo ser suscetível, pode ser modelado por

$$1 - p_{0_i} = \pi(\boldsymbol{\omega}_i) = \frac{1}{1+\exp(\boldsymbol{b}^\top \boldsymbol{\omega}_i)}. \tag{4.12}$$

Outras funções de ligação, tais como as funções probito e complemento log-log, também podem ser consideradas para modelar a probabilidade de ser suscetível, no entanto não serão consideradas neste trabalho.

Para o modelo em (iii), assim como apresentado para os modelos em (i) e (ii), a probabilidade de ser não suscetível pode depender de características individuais, representada por um vetor de covariáveis $\boldsymbol{\omega}_i$. Assim, a probabilidade de um indivíduo ser não suscetível, $p_{0_i} = \exp(-\pi(\boldsymbol{\omega}_i))$, pode ser

modelada como na equação (4.11). Com isso, conclui-se que as covariáveis podem ser introduzidas no parâmetro π por meio da relação

$$\pi(\boldsymbol{\omega}_i) = \log(1 + \exp(\boldsymbol{b}^\top \boldsymbol{\omega}_i)) - \boldsymbol{b}^\top \boldsymbol{\omega}_i. \tag{4.13}$$

Note que os coeficientes $\boldsymbol{\beta}$ (associados à sobrevivência dos indivíduos recorrentes) e $\boldsymbol{b}$ têm interpretações diferentes. Um valor positivo de $\boldsymbol{\beta}$ significa que o risco de experimentar um evento (se suscetível) é maior, enquanto um valor positivo de $\boldsymbol{b}$ significa que a probabilidade de ser não suscetível é maior (ou a probabilidade de ser suscetível é menor). Nos modelos apresentados acima, $(i), (ii)$ e (iii), a probabilidade de ser não suscetível p_0 é definida como a probabilidade de um indivíduo de nunca experimentar qualquer recorrência do evento de interesse. Neste caso, a proporção p_0 é estimada apenas pelos tempos até o primeiro evento, os quais podem descrever dados observados ou censurados. Para os casos em que se observa pelo menos uma recorrência do evento tem-se $Z > 0$, e neste caso, como o indivíduo experimenta recorrência, então o mesmo é suscetível. Por outro lado, para os casos não observados devido a censura à direita, ou seja, não observa-se recorrência, podemos ter $Z = 0$ ou $Z > 0$ e, então, a condição (suscetível ou não suscetível) do indivíduo é desconhecida.

4.1.2 Função de verossimilhança

Sejam n indivíduos sujeitos à ocorrência de um certo evento recorrente. Os dados referentes ao i-ésimo indivíduo, $i = 1, \cdots, n$, são compostos por M_i, que denota os episódios de recorrência no período em estudo; $0 < T_{i1} < \cdots < T_{iM_i}$ os tempos de recorrência do evento; e $Y_{ij} = T_{ij} - T_{i,j-1},\ j = 1, \cdots, M_i$ com $T_{i0} = 0$, que denota os tempos entre eventos. Seja κ_{ij} uma variável indicadora de falha, tal que $\kappa_{ij} = 0$ denota que o tempo do j-ésimo evento é censurado. Então, o número total de recorrências para um indivíduo i é $D_i = \sum_{j=1}^{M_i} \kappa_{ij}$. Assume-se ainda que o tempo de acompanhamento de

um indivíduo em estudo está sujeito a censura à direita, como discutido na Seção 3.1.1. Assim, para o indivíduo i, o tempo de acompanhamento é dado por $\tau_i = \min(\tau, C_i)$, em que τ é o tempo final de estudo e C_i é um tempo de censura à direita não informativo. Sejam $\mathbf{x}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{iM_i})$, com $\mathbf{x}_{ij} = (x_{1ij}, \cdots, x_{pij})$, $j = 1, \cdots, M_i$, um vetor de p covariáveis externas (fixas ou dependentes do tempo), associado aos indivíduos recorrentes, e $\boldsymbol{\omega}_i = (\omega_{1i}, \cdots, \omega_{qi})$ um vetor de q covariáveis associado à distribuição da fragilidade por meio de $\pi(\boldsymbol{\omega}_i)$, expresso em (4.12) ou (4.13). Suponha ainda que as fragilidades Z_i, $i = 1, \cdots, n$, são independentes. O conjunto de dados para o i-ésimo indivíduo é então denotado por $\mathcal{D}_i = \{\boldsymbol{y}_i, \boldsymbol{t}_i, \mathbf{x}_i, \boldsymbol{\omega}_i, \boldsymbol{\kappa}_i, \tau_i\}$, em que $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iM_i})$, $\boldsymbol{t}_i = (t_{i1}, \cdots, t_{iM_i})$ e $\boldsymbol{\kappa}_i = (\kappa_{i1}, \cdots, \kappa_{iM_i})$.

Note que, quando $D_i > 0$, então $Z_i > 0$, o indivíduo é recorrente (suscetível). Quando $D_i = 0$ o indivíduo não apresenta recorrência, e então Z_i não é observado. Assim, para um indivíduo que não apresenta recorrência, isto é, $D_i = 0$, a contribuição para a função de verossimilhança é dada pela função de sobrevivência (4.3). Já para indivíduos com $D_i > 0$, a contribuição é dada pela função densidade (4.4). Para cada indivíduo i, seja $S(\boldsymbol{y}_i|\cdot)$ a sua função de sobrevivência total, considerando todas as suas recorrências, e $f(\boldsymbol{y}_i|\cdot)$ a correspondente função densidade, de modo que

$$S(\boldsymbol{y}_i|\cdot) = \prod_{j=1}^{M_i} S(y_{ij}|t_{i,j-1}) = \exp\left\{-\sum_{j=1}^{M_i} \mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{x}_{ij}} \mathcal{K}_{ij}(\delta)\right\} \tag{4.14}$$

e

$$\begin{aligned} f(\boldsymbol{y}_i|\cdot) &= \prod_{j=1}^{M_i} \left[\lambda_i(y_{ij}|t_{i,j-1})\right]^{c_{ij}} S(y_{ij}|t_{i,j-1}) \\ &= \prod_{j=1}^{M_i} \left[\delta(y_{ij} + t_{i,j-1})^{\delta-1} \mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{x}_{ij}}\right]^{\kappa_{ij}} \exp\left\{-\mathrm{e}^{\boldsymbol{\beta}^\top \mathbf{x}_{ij}} \mathcal{K}_{ij}(\delta)\right\}, \end{aligned} \tag{4.15}$$

sendo $\lambda_i(y_{ij}|t_{i,j-1})$ e $S(y_{ij}|t_{i,j-1})$ dadas, respectivamente, por (3.1) e (3.4), considerando uma função de taxa basal Weibull, e $\mathcal{K}_{ij}(\delta) = (y_{ij} + t_{i,j-1})^\delta - t_{i,j-1}^\delta$.

Considerando o modelo em (i), em que a distribuição de fragilidade segue distribuição Bernoulli, um indivíduo i pode ser tanto um indivíduo recorrente (suscetível) com probabilidade $\pi(\boldsymbol{\omega}_i)$ ou não suscetível com probabilidade $1-\pi(\boldsymbol{\omega}_i)$. Assim, sua contribuição para a função de verossimilhança é dada por

$$1-\pi(\boldsymbol{\omega}_i)+\pi(\boldsymbol{\omega}_i)S(\boldsymbol{y}_i|\cdot). \tag{4.16}$$

Similar a função de verossimilhança para os modelos com fração de cura, a verossimilhança (condicional) dos indivíduos suscetíveis (com $D_i>0$) é dada por

$$\pi(\boldsymbol{\omega}_i)f(\boldsymbol{y}_i|\cdot). \tag{4.17}$$

Combinando as expressões (4.16) e (4.17), a função de verossimilhança conjunta do indivíduo i é dada por

$$\mathcal{L}_i(\boldsymbol{\theta}|\mathcal{D}i)=\Big[\pi(\boldsymbol{\omega}_i)f(\boldsymbol{y}_i|\cdot)\Big]^{d_i}\times\Big[1-\pi(\boldsymbol{\omega}_i)+\pi(\boldsymbol{\omega}_i)S(\boldsymbol{y}_i|\cdot)\Big]^{1-d_i}, \tag{4.18}$$

em que $d_i=I(D_i>0)$, sendo $I(\cdot)$ a função indicadora, $S(\boldsymbol{y}_i|\cdot)$ e $f(\boldsymbol{y}_i|\cdot)$ são as funções de sobrevivência e densidade totais dos indivíduos recorrentes dadas, respectivamente, pelas expressões (4.14) e (4.15), e $\boldsymbol{\theta}$ denota o vetor de parâmetros de interesse.

Analogamente, para o modelo em (ii), em que a distribuição da fragilidade é Geométrica, segue de (4.7) e (4.8) que a função de verossimilhança conjunta do indivíduo i é dada por

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}|\mathcal{D}i) &= \Big[(1-\pi(\boldsymbol{\omega}_i))\pi(\boldsymbol{\omega}_i)f(\boldsymbol{y}_i|\cdot)\{1-\pi(\boldsymbol{\omega}_i)S(\boldsymbol{y}_i|\cdot)\}^{-2}\Big]^{d_i}\\ &\times \Big[(1-\pi(\boldsymbol{\omega}_i))\{1-\pi(\boldsymbol{\omega}_i)S(\boldsymbol{y}_i|\cdot)\}^{-1}\Big]^{1-d_i}.\end{aligned} \tag{4.19}$$

Por fim, para o modelo em (iii), em que a distribuição da fragilidade é Poisson, segue de (4.9) e (4.10) que a função de verossimilhança conjunta do indivíduo i é expressa como

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}|\mathcal{D}i) &= \Big[\pi(\boldsymbol{\omega}_i)f(\boldsymbol{y}_i|\cdot)\exp\{-\pi(\boldsymbol{\omega}_i)(1-S(\boldsymbol{y}_i|\cdot))\}\Big]^{d_i}\\ &\times \Big[\exp\{-\pi(\boldsymbol{\omega}_i)(1-S(\boldsymbol{y}_i|\cdot))\}\Big]^{1-d_i}.\end{aligned} \tag{4.20}$$

Portanto, a função de verossimilhança completa, considerando todos os indivíduos observados, é dada por

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}|\mathcal{D}_i), \tag{4.21}$$

em que $\mathcal{D} = (\mathcal{D}_1, \cdots, \mathcal{D}_n)$ é o conjunto de dados observados considerando os n indivíduos.

4.2 Inferência

Para inferir sobre os parâmetros do modelo adotamos, a princípio, uma abordagem clássica. O objetivo é estimar, além dos parâmetros de regressão $\boldsymbol{\beta}$ associados à distribuição dos indivíduos recorrentes, a proporção de indivíduos não susceptíveis, p_{0_i} e os correspondentes parâmetros de regressão $\boldsymbol{b}$. O procedimento de estimação para os parâmetros do modelo com fragilidade discreta é abordado com base no método de máxima verossimilhança. Intervalos de confiança são construídos para os parâmetros do modelo baseados na aproximação Normal. Como uma alternativa para os casos em que esta aproximação pode não ser válida, um procedimento *bootstrap* não-paramétrico é realizado, considerando B réplicas, obtendo assim os intervalos de confiança *bootstrap*.

4.2.1 Estimação pelo método de máxima verossimilhança

Para construção das equações de verossimilhança e estimação dos parâmetros dos modelos, consideramos uma função de taxa basal Weibull dada por $\lambda_0(y+t) = \alpha\delta(y+t)^{\delta-1}$, de tal forma que a função de taxa $\lambda_i(y_{ij}|t_{i,j-1}, \mathbf{x}_{ij})$ é dada como em (3.11). Isto é,

$$\lambda_i(y_{ij}|t_{i,j-1}, \mathbf{x}_{ij}) = \delta(y_{ij} + t_{i,j-1})^{\delta-1} \exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij}), \qquad \delta > 0, \tag{4.22}$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^\top$ e $\mathbf{x}_{ij} = (1, x_{1ij}, \cdots, x_{pij})$.

Considerando a informação dos n indivíduos, a função de log-verossimilhança baseada no conjunto de dados observado para cada um dos modelos com fragilidade Bernoulli, Geométrica e Poisson é dada, respectivamente, pelas expressões (4.23), (4.24) e (4.25).

• **Modelo com fragilidade Bernoulli**

$$\begin{aligned}\ell(\boldsymbol{\theta}|\mathcal{D}) = \sum_{i=1}^{n} d_i \Bigg\{ & \log(\pi(\boldsymbol{\omega}_i)) + D_i \log(\delta) + \sum_{j=1}^{M_i} \Big[\kappa_{ij}\Big((\delta-1)\log(\tilde{y}_{ij}+t_{i,j-1}) \\ & + \boldsymbol{\beta}^\top \mathbf{x}_{ij}\Big) - \mathcal{K}_{ij}(\delta)\exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij})\Big]\Bigg\} + (1-d_i)\log\Bigg(1-\pi(\boldsymbol{\omega}_i) \\ & + \pi(\boldsymbol{\omega}_i)\exp\Big\{-\sum_{j=1}^{M_i}\mathcal{K}_{ij}(\delta)\exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij})\Big\}\Bigg). \end{aligned} \tag{4.23}$$

• **Modelo com fragilidade Geométrica**

$$\begin{aligned}\ell(\boldsymbol{\theta}|\mathcal{D}) = \sum_{i=1}^{n} d_i \Bigg\{ & \log(1-\pi(\boldsymbol{\omega}_i)) + \log(\pi(\boldsymbol{\omega}_i)) + D_i\log(\delta) \\ & + \sum_{j=1}^{M_i}\Big[\kappa_{ij}\Big((\delta-1)\log(\tilde{y}_{ij}+t_{i,j-1}) + \boldsymbol{\beta}^\top \mathbf{x}_{ij}\Big) - \mathcal{K}_{ij}(\delta)\exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij})\Big] \\ & - 2\log\Bigg(1-\pi(\boldsymbol{\omega}_i)\exp\Big\{-\sum_{j=1}^{M_i}\mathcal{K}_{ij}(\delta)\exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij})\Big\}\Bigg)\Bigg\} \\ & + (1-d_i)\Bigg\{\log(1-\pi(\boldsymbol{\omega}_i)) - \log\Bigg(1-\pi(\boldsymbol{\omega}_i)\exp\Big\{ \\ & - \sum_{j=1}^{M_i}\mathcal{K}_{ij}(\delta)\exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij})\Big\}\Bigg)\Bigg\}. \end{aligned} \tag{4.24}$$

• **Modelo com fragilidade Poisson**

$$\ell(\boldsymbol{\theta}|\mathcal{D}) = \sum_{i=1}^{n} d_i \Bigg\{ \log(\pi(\boldsymbol{\omega}_i)) + D_i \log(\delta) + \sum_{j=1}^{M_i} \Big[\kappa_{ij} \Big((\delta - 1) \log(\tilde{y}_{ij} + t_{i,j-1}) + \boldsymbol{\beta}^\top \mathbf{x}_{ij} \Big) - \mathcal{K}_{ij}(\delta) \exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij}) \Big] \Bigg\} - \pi(\boldsymbol{\omega}_i) \Bigg(1 - \exp \Big\{ - \sum_{j=1}^{M_i} \mathcal{K}_{ij}(\delta) \exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij}) \Big\} \Bigg), \tag{4.25}$$

em que $\tilde{y}_{ij}$ é o valor observado de $\tilde{Y}_{ij} = \min(Y_{ij}, \tau_i - T_{i,j-1})$, $\mathcal{K}_{ij}(\delta)$ é dado por $\mathcal{K}_{ij}(\delta) = (\tilde{y}_{ij} + t_{i,j-1})^\delta - t_{i,j-1}^\delta$ e $\boldsymbol{\theta} = (\delta, \beta_0, \cdots, \beta_p, b_0, \cdots, b_q)$ é o vetor de parâmetros de interesse.

Os estimadores de máxima verossimilhança de $\boldsymbol{\theta}$ podem ser obtidos pela maximização direta da função de log-verossimilhança $\ell(\boldsymbol{\theta}|\mathcal{D})$, utilizando um procedimento de otimização BFGS. Como já discutido na Seção 3.2, inferências sobre os parâmetros do modelo podem ser baseadas, a princípio, nos EMVs e seus erros padrão estimados. Erros padrão assintóticos são obtidos invertendo-se a matriz de informação observada. Intervalos de confiança para os parâmetros podem ser construídos utilizando a aproximação Normal. Para a construção dos intervalos de confiança baseado no método *bootstrap*, os múltiplos tempos de um mesmo indivíduo são selecionados simultaneamente e os intervalos de confiança são obtidos seguindo o procedimento descrito na Seção 3.2.1.

4.3 Estudo de simulação

Nesta seção um estudo de simulação foi realizado com o principal objetivo de investigar algumas propriedades frequentistas do procedimento de estimação dos parâmetros do modelo com fragilidade discreta. O estudo consiste em gerar conjuntos de dados de tamanhos $n = 200$ e 500 do modelo com fragilidade Geométrica (4.7). Por simplicidade, duas covariáveis fixas

são consideradas, sendo x_i associado à distribuição dos indivíduos recorrentes e ω_i, $i = 1, 2, \cdots, n$, associada à função logística, que é considerado para a modelagem da probabilidade de indivíduos não susceptíveis. Os valores das covariáveis x_i e ω_i são ambos gerados de uma distribuição Bernoulli com probabilidade $0,5$. O tempo final de estudo é fixado em $\tau = 9,5$ anos e, para cada indivíduo (ou unidade) i, o tempo de censura C_i é simulado como uma realização independente de uma distribuição Uniforme $U(0, 10)$. Os dados de sobrevivência recorrentes são simulados como segue.

1. Gere um tempo de acompanhamento C_i para cada indivíduo de uma distribuição Uniforme $U(0, 10)$.

2. Faça $\tau_i = \min(\tau, C_i)$.

3. Gere uma variável z_i, indicando quando o indivíduo i é suscetível ou não, de acordo com $z_i \sim$ Geométrica$(1 - \pi(\omega_i))$ com $\pi(\omega_i) = \{1 + \exp(b_0 + b_1\omega_i)\}^{-1}$. Se o indivíduo é não suscetível ($z_i = 0$), então o tempo de recorrência é censurado em τ_i. Se o indivíduo é suscetível ($z_i > 0$), os tempos entre eventos y_{ij} são, então, gerados a partir do modelo (4.1) com função de sobrevivência condicional dada por

$$S(y_{ij}|z_i, t_{i,j-1}) = \exp\left\{-z_i \mathrm{e}^{\beta_0+\beta_1 x_i}\left[(y_{ij} + t_{i,j-1})^{\delta} - t_{i,j-1}^{\delta}\right]\right\},$$

até que $\sum_{j=1}^{M_i} y_{ij} \geq \tau_i$, isto é, para os M_i tempos entre eventos do indivíduo i.

Os parâmetros de interesse são $\boldsymbol{\beta}^\top = (\beta_0, \beta_1)$, $\boldsymbol{b}^\top = (b_0, b_1)$ e δ. Neste estudo de simulação, são atribuídos os seguintes valores aos vetores de parâmetros: $\boldsymbol{\beta}^\top = (1,5; -0,5)$, $\boldsymbol{b}^\top = (-1,0; 0,5)$ e $\delta = 0,6$. As estimativas de máxima verossimilhança (EMVs) são obtidas pela maximização direta da função de log-verossimilhança (4.24) utilizando um procedimento BFGS, mais especificamente a rotina `optim` do *software* R. Para a construção dos intervalos de confiança *bootstrap* foram consideradas $B = 399$

réplicas para cada amostra simulada. Para avaliar o desempenho do estimador de cada parâmetro utilizamos o desvio-padrão (DP) e o vício relativo ($\mathcal{B}$). Os resultados deste estudo de simulação estão resumidos nas Tabelas 4.2 e 4.3. Os resultados para as demais distribuições foram semelhantes a estes apresentados.

A Tabela 4.2 apresenta os valores médios das estimativas pontuais de máxima verossimilhança de cada parâmetro e as probabilidades de cobertura (PC) dos intervalos (assintótico e *bootstrap*) com 95% de confiança. Ao analisar esta tabela, notamos que o método assintótico produz probabilidades de cobertura maiores do que os preditos teoricamente para os parâmetros β_0 e β_1. Já para os parâmetros δ, b_0 e b_1, os resultados das probabilidades de cobertura assintóticos e *bootstrap* não apresentam grandes diferenças, apesar dos resultados obtidos com o método de reamostragem *bootstrap* estarem mais próximos do esperado teoricamente. No geral, as estimativas pontuais dos parâmetros são próximas dos seus verdadeiros valores. O número médio de eventos observado baseado nas 1.000 simulações foi de $6,21$ para ambos os tamanhos de amostra ($n = 200$ e $n = 500$).

Tabela 4.2: Médias das estimativas de máxima verossimilhança e probabilidades de cobertura para os parâmetros do modelo com fragilidade Geométrica.

n	Parâmetro	Média	PC Assint.	PC Boot.
200	β_0	$1,550$	$0,999$	$0,910$
	β_1	-$0,488$	$0,999$	$0,932$
	δ	$0,598$	$0,914$	$0,938$
	b_0	-$0,989$	$0,964$	$0,958$
	b_1	$0,509$	$0,942$	$0,948$
500	β_0	$1,523$	$0,989$	$0,935$
	β_1	-$0,493$	$0,990$	$0,943$
	δ	$0,598$	$0,933$	$0,954$
	b_0	-$0,978$	$0,948$	$0,952$
	b_1	$0,506$	$0,944$	$0,958$

Na Tabela 4.3 apresentamos o desvio padrão empírico das estimativas (DP), a média dos desvios padrão estimados ($\widehat{\text{DP}}$) e a média dos vícios relativos ($\mathcal{B}$). A precisão e relevância da aproximação da variância e covariância das EMVs determinadas a partir da matriz de informação observada é verificada comparando-se com os valores calculados empiricamente. Pode-se notar que os desvios padrão calculados a partir da matriz de informação observada aproximam-se dos desvios padrão empíricos quando o tamanho do conjunto de dados aumenta. Além disso, os desvios padrão das estimativas decrescem quando o número de observações na amostra aumenta, e os vícios aproximam-se de zero para conjuntos de dados de tamanho grande.

Tabela 4.3: Desvio padrão empírico das 1.000 EMVs, média dos desvios padrão estimados e vício médio para o modelo com fragilidade Geométrica.

n	Parâmetro	DP	Média($\hat{\text{DP}}$)	Média($\mathcal{B}$)
200	β_0	$0,445$	$0,482$	$0,122$
	β_1	$0,099$	$0,113$	-$0,025$
	δ	$0,017$	$0,016$	-$0,003$
	b_0	$0,229$	$0,232$	-$0,022$
	b_1	$0,225$	$0,215$	$0,018$
500	β_0	$0,250$	$0,229$	$0,037$
	β_1	$0,044$	$0,037$	-$0,014$
	δ	$0,011$	$0,010$	-$0,004$
	b_0	$0,143$	$0,146$	-$0,011$
	b_1	$0,189$	$0,177$	$0,013$

4.4 Análise de dados reais

Nesta seção apresentamos os resultados da aplicação do modelo proposto para a análise de dados referentes à sucessivas reinternações hospitalares de pacientes diagnosticados com câncer colorretal descrito na Seção 3.4.

O câncer colorretal, se detectado no estágio inicial, tem altas taxas de sobrevivência e cura. De acordo com o Instituto Nacional de Câncer (http://www.cancer.gov/cancertopics/types/colorectal), o câncer colorretal é uma doença altamente tratável e frequentemente curável quando localizada (sem extensão para outros órgãos). A cirurgia é a principal forma de tratamento e, geralmente, resulta em uma taxa de cura de aproximadamente 50% para o câncer de cólon e 45% para o câncer retal (Yu, 2008). Sendo assim, é razoável assumir que uma proporção de indivíduos são curados (deixam de ser susceptíveis) e não apresentam recorrência no futuro.

O Modelo II foi então ajustado aos dados de reinternação hospitalar abordando as quatro covariáveis descritas anteriormente, sendo elas: quimioterapia (x_1, em que 1: recebeu quimioterapia e 0: caso contrário); sexo (x_2, em que 1: feminino e 0: masculino); estágio do tumor, de acordo com a classificação de Dukes, (x_3, em que 1: estágio A-B, 2: estágio C e 3: estágio D); e índice de comorbidade de Charlson (x_4, em que 0: índice 0, 1: índice $1-2$ e 3: índice ≥ 3). Para cada indivíduo, os vetores de covariáveis associados são $\mathbf{x}_{ij} = (x_{1i}, x_{2i}, x_{3i}^*, x_{4ij}^*)^\top$, com $x_{3i}^* = (I(x_{3i}=2), I(x_{3i}=3))$ e $x_{4ij}^* = (I(x_{4ij}=1), I(x_{4ij}=3))$, e $\boldsymbol{\omega}_i = (x_{1i}, x_{2i}, x_{3i}^*)^\top$. Para a fragilidade são consideradas as três distribuições: Bernoulli, Geométrica e Poisson. A Tabela 4.4 apresenta os valores de máximo da log-verossimilhança, max $\ell(\cdot)$, e do critério de informação de Akaike, AIC (*Akaike Information Criterion*), para os três modelos ajustados. Comparando estes critérios, notamos evidência a favor do modelo com fragilidade Bernoulli, o qual apresenta o maior valor para o critério max $\ell(\cdot)$ e consequentemente o menor AIC.

Os resultados das estimativas de máxima verossimilhança dos parâmetros do modelo com fragilidade Bernoulli, seus erros padrão (DP), bem como seus intervalos *bootstrap* com 95% de confiança ($\text{IC}_{95\%}$) e valor-p são apresentados na Tabela 4.5. Esta tabela mostra que apenas o estágio do tumor com classificação D está significativamente relacionado com a taxa de cura

Tabela 4.4: Valores de máximo da log-verossimilhança e critério AIC para os três modelos ajustados aos dados de reinternação hospitalar.

	Distribuições para a fragilidade		
Critério	Bernoulli	Geométrica	Poisson
max $\ell(\cdot)$	-3371, 661	-3484, 523	-3439, 515
AIC	6769, 322	6995, 046	6905, 029

(proporção de indivíduos não susceptíveis) e, para os indivíduos não curados, os principais fatores de risco para reinternação devido à recorrência de câncer colorretal são sexo, estágio do tumor com classificação D e índice de Charlson. Na Tabela 4.5, o símbolo $*$ indica que o valor-p é não significativo ao nível de 5%.

Os resultados do ajuste que considera apenas as covariáveis significativas ao nível de 5% são apresentados nas Tabelas 4.6 e 4.7 e, novamente, dão evidências a favor do modelo com fragilidade Bernoulli.

Os resultados da Tabela 4.7 mostram que os tempos entre reinternações são menores para pacientes com estágio do tumor D, ou seja, esta covariável contribui para o aumento do risco de reinternação devido à recorrência de câncer colorretal, com risco relativo igual a $1,89$ ($\text{IC}_{95\%} = (1,49;\ 2,39)$). Além disso, para estes pacientes, a probabilidade de cura diminui consideravelmente quando comparado com o grupo de referência.

Usando o modelo sem fragilidade (Modelo I), concluímos que pacientes do sexo masculino têm risco significativamente maior de apresentar recorrências do que os pacientes do sexo feminino, com risco relativo igual a $1,71$. O modelo com fragilidade Bernoulli fornece maiores detalhes, isto é, o sexo não influencia significativamente na probabilidade de cura e, pacientes do sexo masculino não

Tabela 4.5: Estimativas dos parâmetros do modelo com fragilidade Bernoulli para os dados de reinternação hospitalar.

Descrição	Parâmetro	Estimativa	DP	$IC_{95\%}$	valor-p
Sobrev. recorrentes					
Intercepto	β_0	-5,190	0,281	(-5,742; -4,638)	$< 0,001$
Quimio	β_1	-0,095	0,122	(-0,334; 0,144)	$0,451^*$
Sexo (Fem.)	β_2	-0,543	0,118	(-0,775; -0,311)	$< 0,001$
Estágio Tumor					
A-B (ref)					
C	β_{3_1}	0,288	0,141	(-0,011; 0,565)	$0,064^*$
D	β_{3_2}	0,872	0,155	(0,568; 1,176)	$< 0,001$
Índice Charlson					
0 (ref)					
$1-2$	β_{4_1}	0,592	0,219	(0,163; 1,022)	0,019
≥ 3	β_{4_2}	0,754	0,123	(0,513; 0,995)	$< 0,001$
Parâmetro					
Weibull	δ	0,810	0,034	(0,743; 0,877)	$< 0,001$
Modelo logístico					
Intercepto	b_0	-0,531	0,327	(-1,172; 0,111)	$0,131^*$
Quimio	b_1	0,285	0,325	(-0,352; 0,921)	$0,398^*$
Sexo (Fem.)	b_2	-0,028	0,313	(-0,642; 0,585)	$0,929^*$
Estágio Tumor					
A-B (ref)					
C	b_{3_1}	-0,113	0,330	(-0,759; 0,534)	$0,739^*$
D	b_{3_2}	-1,643	0,679	(-2,974; -0,312)	0,032

Tabela 4.6: Valores de máximo da log-verossimilhança e critério AIC para os três modelos ajustados aos dados de reinternação hospitalar, considerando apenas as covariáveis significativas.

	Distribuições para a fragilidade		
Critério	Bernoulli	Geométrica	Poisson
max $\ell(\cdot)$	-3381, 005	-3496, 202	-3443, 660
AIC	6776, 009	7006, 403	6901, 320

Tabela 4.7: Estimativas dos parâmetros do modelo com fragilidade Bernoulli para os dados de reinternação hospitalar considerando apenas as covariáveis significativas.

Descrição	Parâmetro	Estimativa	DP	$IC_{95\%}$	valor-p
Sobrev. recorrentes					
Intercepto	β_0	-5, 032	0, 254	(-5, 529; -4, 535)	$< 0,001$
Sexo (Fem.)	β_2	-0, 492	0, 108	(-0, 704; -0, 279)	0, 004
Estágio Tumor					
A-B (ref)					
D	β_{3_2}	0, 636	0, 120	(0, 400; 0, 872)	0, 002
Índice Charlson					
0 (ref)					
$1 - 2$	β_{4_1}	0, 554	0, 214	(0, 134; 0, 974)	0, 041
≥ 3	β_{4_2}	0, 797	0, 119	(0, 565; 1, 030)	$< 0,001$
Parâmetro					
Weibull	δ	0, 809	0, 034	(0, 743; 0, 876)	$< 0,001$
Modelo logístico					
Estágio Tumor					
A-B (ref)					
D	b_{3_2}	-2, 120	0, 643	(-3, 381; -0, 859)	0, 017

curados (susceptíveis) têm risco maior de recorrência, com risco relativo igual a $1,64$ ($\text{IC}_{95\%} = (1,36;\ 2,02)$). Ainda, para os pacientes recorrentes, o índice de comorbidade de Charlson contribui significativamente para o aumento do risco de reinternação, com risco relativo de $1,74$ ($\text{IC}_{95\%} = (1,14;\ 2,65)$) para índice de Charlson $1-2$, e $2,22$ ($\text{IC}_{95\%} = (1,76;\ 2,80)$) para índice de Charlson ≥ 3. Estes resultados demonstram o impacto da proporção de indivíduos não susceptíveis no modelo e, novamente, fornecem conclusões que estão em concordância com as conclusões obtidas de pesquisas que utilizaram o mesmo conjunto de dados em um contexto de eventos recorrentes, como por exemplo Yu (2008) e Rondeau *et al.* (2011).

4.5 Alguns comentários

Neste capítulo propomos um modelo para dados multivariados, em particular para dados na presença de eventos recorrentes, induzido por uma fragilidade discreta. Este modelo acomoda dados para eventos recorrentes e estende os modelos de fragilidade existentes, permitindo uma distribuição discreta para a variável que descreve a fragilidade. Dados de sobrevivência em que uma proporção de indivíduos na população deixa de ser suscetível a determinado evento de interesse são encontrados em diversas áreas, inclusive em situações em que a recorrência de eventos existe e não deve ser ignorada. Nesse sentido, o modelo proposto tem flexibilidade para incluir estas situações em que existe a possibilidade de indivíduos com risco zero, uma vez que a fragilidade assumir valor zero corresponde a um modelo que contém uma fração de indivíduos que nunca falham. As estimativas dos parâmetros são obtidas utilizando um procedimento de máxima verossimilhança. Os resultados da simulação mostram a eficácia do método de estimação dos parâmetros e, na obtenção das probabilidades de cobertura, mostram um melhor desempenho do método *bootstrap*.

A importância prática do Modelo II proposto foi demonstrada utilizando um conjunto de dados reais referente à hospitalizações de pacientes diagnosticados com câncer colorretal. Os resultados obtidos forneceram maiores detalhes quando comparados ao modelo proposto no Capítulo 3, em que todos os indivíduos são tratados como recorrentes.

Capítulo 5

Aplicação aos Dados de Malária

Nesta seção apresentamos uma aplicação dos modelos abordados nos Capítulos 3 e 4 a um banco de dados real envolvendo recorrências de malária em indivíduos atendidos pela Faculdade de Medicina da UFMT.

5.1 Apresentação do banco de dados

O banco de dados apresentado nesta seção refere-se à recaídas de malária em indivíduos atendidos pela Faculdade de Medicina da Universidade Federal de Mato Grosso (UFMT), Cuiabá, Brasil, o qual foi gentilmente fornecido pelo professor Doutor Cor Jesus Fernandes Fontes.

A malária é uma doença parasitária de grande importância e prevalência da atualiadade, constituindo ainda um problema de saúde pública. Essa doença ocorre principalmente em países com clima tropical e subtropical, sendo causada por protozoários parasitas do gênero *Plasmodium* que atacam as células vermelhas do sangue. Existem mais de 100 diferentes espécies de *Plasmodium*, sendo cinco delas que comumente infectam os seres humanos: *P. falciparum*, *P. malariae*, *P. vivax*, *P. ovale* e *P. knowlesi*. Apenas as três primeiras são prevalentes no Brasil. Segundo Kirchgatter & del Portillo (1998), as recaídas de malária são definidas como as novas manifestações de

uma infecção provocada por um parasita. As recaídas podem ser classificadas como precoces, ocorrendo antes de 2 meses após o ataque primário, ou tardias, ocorrendo após 6 meses do ataque primário, sendo o ataque primário definido como o ataque malárico que marca o fim do período de incubação (primeiros sintomas). Muitos autores têm realizado pesquisas em relação à frequência das recaídas (Boulos *et al.*, 1991; Bunnag *et al.*, 1994), e relatam que as recaídas ocorrem com maior frequência no período de 6 meses após o tratamento.

Os dados coletados são referentes ao ataque primário e à recaída de malária em pacientes atendidos na UFMT. Neste estudo, foram incluídos somente pacientes que afirmaram ter permanecido fora de áreas com risco de transmissão de malária após o ataque primário e ter feito uso correto da medicação, sem apresentar intercorrências. Cada paciente foi acompanhado por um período de 1 ano. As datas em que os indivíduos deram entrada no hospital com os primeiros sintomas de malária foram registradas, considerando-se como dia 0 o dia do diagnóstico. Estas datas foram convertidas no número de dias entre as sucessivas ocorrências da doença, compondo assim o conjunto de dados. O número médio de recorrências observado por indivíduo é de $0,487$ (variando de 0 a 6). Cada paciente, após o diagnóstico inicial, recebeu um dos três tratamentos disponíveis: cloroquina+primaquina (CR+PR), artemeter+lumefantrina (AR+LF) ou artemeter+lumefantrina+primaquina (AR+LF+PR), de acordo com o resultado para a espécie do parasita causador da malária. Informações adicionais sobre os pacientes, tais como sexo e idade, também foram incluídas. Assim as covariáveis que compõem o conjunto de dados são: idade (x_1, em que 1 : maior ou igual a 37 anos e 0 : menor que 37 anos); sexo (x_2, em que 1 : masculino e 0 : feminino); e tratamento (x_3, em que 1 : recebeu CR+PR, 2 : recebeu AR+LF e 3 : recebeu AR+LF+PR). A porcentagem de dados censurados é $60,8\%$.

5.2 Análise dos dados

Dada a relevância de casos de malária no Brasil, faz-se necessário o estudo de ferramentas estatísticas apropriadas que auxiliem na interpretação dos parâmetros dos modelos e que podem ser aplicadas na análise dos dados relacionados ao tratamento dessa doença. Com isso, as metodologias propostas são aplicadas ao dados de malária no intuito de determinar quais covariáveis estão relacionadas com um aumento e/ou diminuição do tempo até a recorrência do evento de interesse e, também, estimar a probabilidade de indivíduos não suscetíveis por meio da introdução de uma fragilidade discreta no modelo.

Os Modelos I e II são utilizados para a análise do conjunto de dados referente ao estudo de malária descrito na Seção 5.1. Os efeitos das covariáveis mencionadas foram avaliados tanto na proporção de indivíduos não suscetíveis quanto na sobrevivência dos indivíduos recorrentes. A Tabela 5.1 apresenta os valores de máximo da log-verossimilhança, max $\ell(\cdot)$, e o valor da estatística AIC para os três modelos com fragilidade ajustados. Comparando essas estatísticas, notamos que as diferença são pequenas entre os modelos, embora evidenciam a favor do modelo com fragilidade Poisson. Vale ressaltar que os valores das estatísticas max $\ell(\cdot)$ e AIC do modelo com fragilidade Geométrica, na comparação com os demais modelos, o classifica como o menos adequado.

Tabela 5.1: Valores de máximo da log-verossimilhança e critério AIC para os três modelos com fragilidde ajustados aos dados de malária.

	Distribuições para a fragilidade		
Critério	Bernoulli	Geométrica	Poisson
max $\ell(\cdot)$	-336, 16	-337, 11	-335, 15
AIC	698, 32	700, 22	696, 30

Os resultados das estimativas de máxima verossimilhança dos parâmetros do modelo com fragilidade Poisson e seus desvios padrão (DP), bem como os resultados para o modelo sem fragilidade (Modelo I) com função de taxa basal Weibull, abordado no Capítulo 3, são apresentados na Tabela 5.2. O símbolo $*$ indica que o valor-p é não significativo ao nível de 5%. Com base no valor de máximo da log-verossimilhança, apresentado na Tabela 5.2, o modelo com fragilidade Poisson fornece um melhor ajuste quando comparado ao modelo sem fragilidade.

Para os indivíduos recorrentes, baseado no modelo com fragilidade Poisson, o valor positivo de $\hat{\beta}_1$ indica que os tempos entre os sucessivos episódios de malária para pacientes com idade maior ou igual a 37 anos são significativamente menores do que para pacientes com idade inferior a 37 anos. Com isso, pode-se observar que um paciente com idade maior ou igual a 37 anos tem um risco 40% maior de apresentar recorrências do que um paciente com idade inferior. As covariáveis sexo (masculino) e tratamento estão associadas a tempos entre recorrências maiores. Dessa forma, estas covariáveis contribuem para a diminuição do risco de recorrência de malária, com risco relativo igual a $0,77$ ($\text{IC}_{95\%} = (0,61;\ 0,97)$) para pacientes do sexo masculino, e riscos relativos iguais a $0,54$ ($\text{IC}_{95\%} = (0,29;\ 0,98)$) e $0,34$ ($\text{IC}_{95\%} = (0,22;\ 0,52)$) para tratamentos com AR+LF e AR+LF+PR, respectivamente. Por outro lado, se a proporção de indivíduos não suscetíveis (ou curados, no caso da malária) for ignorada, as covariáveis idade, sexo e tratamento com AR+LF não apresentam efeito significativo sobre os tempos de recorrência de malária. Isto pode ser observado a partir dos resultados na Tabela 5.2 (coluna 3), e mostra o impacto da proporção de indivíduos curados (não suscetíveis) na modelagem.

Ainda, a partir da equação (4.11) e dos resultados dispostos na Tabela 5.2, a probabilidade de um indivíduo ser curado, isto é, não suscetível à recorrência de malária é de 77% para um indivíduo pertencente ao grupo de referência

Tabela 5.2: Estimativas dos parâmetros do modelo sem fragilidade e do modelo com fragilidade Poisson para os dados de malária.

		Modelo sem frag.		Modelo com frag. Poisson	
Descrição	Parâmetro	Estimativa	DP	Estimativa	DP
Sobrev. recorrentes					
Intercepto	β_0	-4,488	0,545	-5,360	0,238
Idade ($\geq$ 37 anos)	β_1	-0,270*	0,242	0,335	0,124
Sexo (Masc.)	β_2	-0,235*	0,284	-0,259	0,117
Tratamento					
CR+PR (ref)					
AR+LF	β_{3_1}	0,043*	0,274	-0,625	0,311
AR+LF+PR	β_{3_2}	-6,248	0,621	-1,088	0,223
Parâmetro					
Weibull	δ	0,720	0,080	0,758	0,102
Modelo logístico					
Intercepto	b_0	–	–	1,217	0,390
Idade ($\geq$ 37 anos)	b_1	–	–	0,049*	0,144
Sexo (Masc.)	b_2	–	–	0,114	0,070
Tratamento					
CR+PR (ref)					
AR+LF	b_{3_1}	–	–	-0,320	0,104
AR+LF+PR	b_{3_2}	–	–	1,816	0,523
Log-verossimilhança		-534,82		-335,15	

(indivíduos com idade inferior a 37 anos, do sexo feminino e que recebeu tratamento com CR+PR). Para os indivíduos que receberam tratamento com AR+LF a probabilidade de cura é de 71%. Já para os indivíduos que receberam tratamento com AR+LF+PR a probabilidade de cura é de 95%, indicando a eficácia deste tipo de tratamento.

Capítulo 6

Conclusões e Propostas Futuras

As metodologias apresentadas neste trabalho foram construídas para análise e modelagem de dados de eventos recorrentes. O primeiro modelo abordado possibilita investigar tempos entre sucessivas oorrências de um evento de interesse com uma dependência induzida pelo tempo de censura quando o número de recorrências do indivíduo é maior do que 1. Neste modelo, o processo de recorrência de eventos segue um modelo de Poisson não homogêneo e a função de taxa associada é caracterizada por uma estrutura multiplicativa. A distribuição condicional do tempo entre eventos foi derivada facilmente da função de taxa marginal, que é uma formulação atrativa para dados de eventos recorrentes e possibilita interpretações práticas mais diretas. O procedimento inferencial foi baseado na abordagem de máxima verossimilhança, o qual possibilitou fácil implementação sem esforços computacionais. O procedimento de estimação mostrou-se eficaz e as inferências sobre os parâmetros foram bastante satisfatórias, mesmo para uma quantidade pequena de unidades na amostra. Com a constatação do bom desempenho dessa abordagem, consideramos o ajuste do modelo para análise de um conjunto de dados reais referente à sucessivas reinternações de pacientes diagnosticados com câncer colorretal. A metodologia e os resultados obtidos a partir da mesma foram comparados com aqueles de estudos anteriores. Ainda, para este

modelo gráficos dos resíduos de Cox-Snell e martingale foram apresentados e, mostraram-se de grande importância por permitir uma avaliação mais direta do modelo para dados de eventos recorrentes.

Uma extensão dos modelos de fragilidade para dados de eventos recorrentes foi feita considerando-se o uso de distribuições Bernoulli, Geométrica, Poisson ou outra distribuição discreta para a variável de fragilidade. Essa extensão teve como objetivo contemplar indivíduos com risco zero para a recorrência de um determinado evento. Para a construção desse modelo foi utilizado como base o modelo para eventos recorrentes abordado na primeira etapa. Novamente, para o ajuste do modelo com fragilidade discreta utilizamos uma abordagem clássica baseada em máxima verossimilhança. Visando avaliar o desempenho da abordagem clássica, realizamos um estudo de simulação, o qual considerou apenas o modelo com fragilidade Geométrica. O método de máxima verossimilhança resultou em estimativas pontuais satisfatórias. No entanto, as inferências sobre os parâmetros, com base em intervalos de confiança, foram mais precisas quando consideramos o método de reamostragem *bootstrap* do que as obtidas com base na teoria assintótica. Os modelos com fragilidades discretas (Bernoulli, Geométrica e Poisson) foram utilizados para a análise do conjunto de dados referente à sucessivas reinternações de pacientes diagnosticados com câncer colorretal, apontando o modelo com fragilidade Bernoulli como o mais adequado para este conjunto de dados.

Ainda, os modelos tratados neste trabalho foram aplicados a um conjunto de dados reais referente à recaídas de malária em pacientes atendidos pela Faculdade de Medicina da UFMT. Com a análise dos resultados, que apontou o modelo com fragilidade Poisson o mais adequado, identificamos as covariáveis relacionadas ao aumento e/ou diminuição do risco de recorrência de malária para os pacientes suscetíveis e, ainda, estimamos a probabilidade de indivíduos não suscetíveis (indivíduos com risco zero).

Por fim, existem várias pesquisas que podem ser realizadas como conti-

nuação da desenvolvida na presente tese. Dentre estas, propomos os seguintes tópicos:

1. Investigação e obtenção da estrutura de dependência entre os tempos dos eventos para o modelo em (3.1). Obtenção de coeficientes que medem a dependência;

2. Desenvolver procedimentos inferenciais em uma perspectiva Bayesiana para o modelo em (3.1) e um estudo de influência caso a caso na linha de Cho *et al.* (2009);

3. Considerar outras formas alternativas (paramétricas e não paramétricas), na linha de Rondeau *et al.* (2011), para a função de taxa basal dos indivíduos recorrentes no modelo em (4.3) com o intuito de fornecer maior flexibilidade na forma da função de taxa/sobrevivência;

4. Considerar distribuições de fragilidade discretas mais gerais, tais como a distribuição Binomial Negativa, que é estatisticamente conveniente uma vez que inclui outras distribuições como casos especiais (Geométrica) ou no limite (Poisson), e a distribuição Weibull Discreta (Bakouch *et al.*, 2014);

5. Investigar o uso de distribuições de fragilidade discretas em modelos semiparamétricos para eventos recorrentes;

6. Considerar procedimentos inferenciais em uma perspectiva clássica para o modelo em (4.3), utilizando o algoritmo EM (Tanner, 1996);

7. Considerar procedimentos de estimação e diagnóstico em uma perspectiva Bayesiana para o modelo em (4.3), utilizando métodos de Monte Carlo em Cadeia de Markov;

8. Considerar uma probabilidade de se tornar não suscetível após a ocorrência de cada evento, modelando um $\pi(\boldsymbol{\omega}_{ij}) = \pi_{ij}$ ao invés de $\pi(\boldsymbol{\omega}_i) = \pi_i$.

Referências

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, **6**, 701–726.

Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical Statistics and Probability Theory*, pages 1–25. Springer, New York.

Aalen, O. O. & Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine*, **10**, 1227–1240.

Aalen, O. O., Fosen, J., Weedon-Fekjaer, H., Borgan, O. & Husebye, E. (2004). Dynamic analysis of multivariate failure time data. *Biometrics*, **60**, 764–773.

Aalen, O. O., Borgan, O. & Gjessing, H. K. (2008). *Survival and event history analysis: a process point of view*. Springer, New York.

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, **10**, 1100–1120.

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag, New York.

Ata, N. & Özel, G. (2013). Survival functions for the frailty models based on the discrete compound poisson process. *Journal of Statistical Computation and Simulation*, **83**, 2105–2116.

Bakouch, H. S., Jazi, M. A. & Nadarajah, S. (2014). A new discrete distribution. *Statistics*, **48**, 200–240.

Bao, Y., Dai, H., Wang, T. & Chuang, S. K. (2013). A joint modelling approach for clustered recurrent events and death events. *Journal of Applied Statistics*, **40**, 123–140.

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**, 501–515.

Berman, M. & Turner, T. R. (1992). Approximating point process likelihoods with glim. *Applied Statistics*, **41**, 31–38.

Bijwaard, G. E., Franses, P. H. & Paap, R. (2006). Modeling purchases as repeated events. *Journal of Business & Economic Statistics*, **24**, 487–502.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B*, **11**, 15–53.

Borgan, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scandinavian Journal of Statistics*, **11**, 1–16.

Boulos, M., Amato Neto, V., Dutra, A. P., Di Santi, S. M. & Shiroma, M. (1991). Análise da frequência de recaídas de malária por Plasmodium vivax em região não endêmica (São Paulo, Brasil). *Rev. Inst. Med. trop. S. Paulo*, **33**, 143–146.

Box-Steffensmeier, J. M. & De Boef, S. (2006). Repeated events survival models: The conditional frailty model. *Statistics in Medicine*, **25**, 3518–3533.

Bunnag, D., Karbwang, J., Thanavibul, A., Chittamas, S., Ratanapongse, Y., Chalermrut, K., Bangchang, K. N. & Harinasuta, T. (1994). High dose

of primaquine in primaquine resistant vivax malaria. *Trans. R. Soc. Trop. Med. Hyg.*, **88**, 218–219.

Cancho, V. G., Louzada, F. & Barriga, G. D. C. (2012). The geometric Birnbaum-Saunders regression model with cure rate. *Journal of Statistical Planning and Inference*, **142**, 993–1000.

Caroni, C., Crowder, M. & Kimber, A. (2010). Proportional hazards models with discrete frailty. *Lifetime Data Analysis*, **16**, 374–384.

Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Barbosa, M. T. S. & Shimakura, S. E. (2005). *Análise de sobrevida: teoria e aplicações em saúde*. Editora Fiocruz, Rio de Janeiro.

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

Chen, Y. Q., Wang, M. C. & Huang, Y. (2004). Semiparametric regression analysis on longitudinal pattern of recurrent gap times. *Biostatistics*, **5**, 277–290.

Chiang, C. T., James, L. F. & Wang, M. C. (2005). Random weighted bootstrap method for recurrent events with informative censoring. *Lifetime Data Analysis*, **11**, 489–509.

Cho, H., Ibrahim, J. G., Sinha, D. & Zhu, H. (2009). Bayesian case influence diagnostics for survival models. *Biometrics*, **65**, 116–124.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141–151.

Collett, D. (2003). *Modelling survival data in medical research*. Chapman and Hall/CRC, Florida.

Colosimo, E. A. & Giolo, S. R. (2006). *Análise de sobrevivência aplicada*. Edgard Blucher, São Paulo.

Cook, R. J. & Lawless, J. F. (2002). Analysis of repeated events. *Statistical Methods in Medical Research*, **11**, 141–166.

Cook, R. J. & Lawless, J. F. (2007). *The statistical analysis of recurrent events*. Springer-Verlag, New York.

Cox, D. R. (1972a). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, **34**, 187–220.

Cox, D. R. (1972b). The statistical analysis of dependencies in point processes. In *Stochastic Point Processes*, pages 55–66, New York. John Wiley & Sons.

Crowther, M. J. & Lambert, P. C. (2014). A general framework for parametric survival analysis. *Statistics in Medicine*, **33**, 5280–5297.

Dong, L. & Sun, L. (2015). A flexible semiparametric transformation model for recurrent event data. *Lifetime Data Analysis*, **21**, 20–41.

Duchateau, L. & Janssen, P. (2008). *The frailty model*. Springer, New York.

Duchateau, L., Janssen, P., Kezic, I. & Fortpied, C. (2003). Evolution of recurrent asthma event rate over time in frailty models. *Journal of the Royal Statistical Society. Series C*, **52**, 355–363.

Finkelstein, M. (2008). *Failure rate modelling for reliability and risk*. Springer, New York.

Fleming, T. R. & Harrington, D. P. (1991). *Counting processes and survival analysis*. John Wiley & Sons, New York.

Fogo, J. C. (2007). *Modelo de regressão para um processo de renovação Weibull com termo de fragilidade*. Tese (Doutorado em Estatística e Experimentação Agronômica), Universidade de São Paulo, Piracicaba.

Fosen, J., Borgan, O., Weedon-Fekjaer, H. & Aalen, O. O. (2006a). Dynamic analysis of recurrent event data using the additive hazard model. *Biometrical Journal*, **48**, 381–398.

Fosen, J., Ferkingstad, E., Borgan, O. & Aalen, O. O. (2006b). Dynamic path analysis - a new approach to analyzing time-dependent covariates. *Lifetime Data Analysis*, **12**, 143–167.

Fredette, M. & Lawless, J. F. (2007). Finite-horizon prediction of recurrent events, with application to forecasts of warranty claims. *Technometrics*, **49**, 66–80.

Ghosh, D. (2004). Accelerated rates regression models for recurrent failure time data. *Lifetime Data Analysis*, **10**, 247–261.

Gill, R. D. (1984). Understanding Cox's regression model: A martingale approach. *Journal of the American Statistical Association*, **79**, 441–447.

Gill, R. D. & Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, **18**, 1501–1555.

Giolo, S. R. (2003). *Variáveis latentes em análise de sobrevivência e curvas de crescimento*. Tese (Doutorado em Estatística e Experimentação Agronômica), Universidade de São Paulo, Piracicaba.

Gonzalez, J. R., Fernandez, E., Moreno, V., Ribes, J., Peris, M., Navarro, M., Cambray, M. & Borras, J. M. (2005). Sex differences in hospital readmission among colorectal cancer patients. *Journal of Epidemiology and Community Health*, **59**, 506–511.

Gouvêa, G. D. R. (2010). *Métodos bayesianos para análise de dados de eventos recorrentes considerando uma classe geral de modelos com fragilidade multiplicativa*. Tese (Doutorado em Estatística e Experimentação Agropecuária), Universidade Federal de Lavras, Lavras.

Hougaard, P. A. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika*, **71**, 75–83.

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

Hougaard, P. A. (1986b). A class of multivanate failure time distributions. *Biometrika*, **73**, 671–678.

Hougaard, P. A. (2000). *Analysis of multivariate survival data*. Springer, New York.

Huang, Y. & Chen, Y. Q. (2003). Marginal regression of gaps between recurrent events. *Lifetime Data Analysis*, **9**, 293–303.

Ibrahim, J. G., Chen, M. H. & Sinha, D. (2005). *Bayesian survival analysis*. Springer, New York.

Jacod, J. (1975). Multivariate point processes: predictable projection, radon-nikodym derivatives, representation of martingales. *Probability Theory and Related Fields*, **31**, 235–253.

Jin, Z., Lin, D. Y. & Ying, Z. (2006). Rank regression analysis of multivariate failure time data based on marginal linear models. *Scandinavian Journal of Statistics*, **33**, 1–23.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The statistical analysis of failure time data*. John Wiley & Sons, Hoboken, NJ.

Kelly, P. J. & Lim, L. L. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine*, **19**, 13–33.

Kirchgatter, K. & del Portillo, H. A. (1998). Molecular analysis of plasmodium vivax relapses using the msp1 molecule as a genetic marker. *Journal of Infectious Diseases*, **177**, 511–555.

Lawless, J. F. (1987). Regression methods for poisson process data. *Journal of the American Statistical Association*, **82**, 808–815.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. John Wiley & Sons, Hoboken, NJ.

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

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

Lim, H. J. & Zhang, X. (2009). Semi-parametric additive risk models: Application to injury duration study. *Accident Analysis and Prevention*, **41**, 211–216.

Lim, H. J. & Zhang, X. (2011). Additive and multiplicative hazards modeling for recurrent event data analysis. *BMC Medical Research Methodology*, **11**, 1–12.

Lim, H. J., Liu, J. & Melzer-Lange, M. (2007). Comparison of methods for analyzing recurrent events data: application to the emergency department visits of pediatric firearm victims. *Accident Analysis & Prevention*, **39**, 290–299.

Lin, D. Y., Wei, L. J. & Ying, Z. L. (1998). Accelerated failure time models for counting processes. *Biometrika*, **85**, 605–618.

Lin, D. Y., Sun, W. & Ying, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika*, **86**, 59–70.

Lin, D. Y., Wei, L. J., Yang, I. & Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society. Series B*, **62**, 711–730.

Liu, B., Lu, W. & Zhang, J. (2014). Accelerated intensity frailty model for recurrent events data. *Biometrics*, **70**, 579–587.

Louzada, F. & Cobre, J. (2012). A multiple time scale survival model with a cure fraction. *Test*, **21**, 355–368.

Louzada, F., Macera, M. A. C. & Cancho, V. G. (2015). The poisson-exponential model for recurrent event data: an application to bowel motility data. *Journal of Applied Statistics*, **42**, 2353–2366.

Louzada-Neto, F. (2004). A hybrid scale intensity model for recurrent event data. *Communications in Statistics-Theory and Methods*, **33**, 119–133.

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

Luo, X. & Huang, C. Y. (2011). Analysis of recurrent gap time data using the weighted risk-set method and the modified within-cluster resampling method. *Statistics in Medicine*, **30**, 301–311.

Macera, M. A. C., Louzada, F., Cancho, V. G. & Fontes, C. J. F. (2014). The exponential-poisson model for recurrent event data: An application to a set of data on malaria in brazil. *Biometrical Journal*, **57**, 201–214.

Martinussen, T. & Scheike, T. H. (2007). *Dynamic regression models for survival data*. Springer-Verlag, New York.

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

Monaco, J., Cai, J. & Grizzle, J. (2005). Bootstrap analysis of multivariate failure time data. *Statistics in Medicine*, **24**, 3387–3400.

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

Oakes, D. A. (1992). Frailty models for multiple event times. In *Survival analysis: state of the art*, pages 371–379. Springer, New York.

Peng, Y. & Xu, J. (2012). An extended cure model and model selection. *Lifetime Data Analysis*, **18**, 215–233.

Peng, Y. W., Taylor, J. M. G. & Yu, B. B. (2007). A marginal regression model for multivariate failure time data with a surviving fraction. *Lifetime Data Analysis*, **13**, 351–369.

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

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, New York.

Price, D. L. & Manatunga, A. K. (2001). Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine*, **20**, 1515–1527.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rigdon, S. E. & Basu, A. P. (2000). *Statistical methods for the reliability of repairable systems*. John Wiley & Sons, New York.

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

Rodrigues, J., de Castro, M., Balakrishnan, N. & Cancho, V. G. (2011). Destructive weighted poisson cure rate models. *Lifetime Data Analysis*, **17**(3), 333–346.

Rondeau, V., Schaffner, E., Corbière, F., Gonzalez, J. R. & Mathoulin-Pélissier, S. (2011). Cure frailty models for survival data: application to recurrences for breast cancer and to hospital readmissions for colorectal cancer. *Statistical Methods in Medical Research*, **22**, 243–260.

Sankaran, P. G. & Anisha, P. (2011). Shared frailty model for recurrent event data with multiple causes. *Journal of Applied Statistics*, **38**, 2859–2868.

Sankaran, P. G. & Anisha, P. (2012). Additive hazards models for gap time data with multiple causes. *Statistics and Probability Letters*, **82**, 1454–1462.

Santos, D. M., Davies, R. B. & Francis, B. (1995). Nonparametric hazard versus nonparametric frailty distribution in modelling recurrence of breast cancer. *Journal of Statistical Planning and Inference*, **47**, 111–127.

Schaubel, D. E. & Cai, J. W. (2004a). Regression methods for gap time hazard functions of sequentially ordered multivariate failure time data. *Biometrika*, **91**, 291–303.

Schaubel, D. E. & Cai, J. W. (2004b). Non-parametric estimation of gap time survival functions for ordered multivariate failure time data. *Statistics in Medicine*, **23**, 1885–1900.

Schaubel, D. E., Zeng, D. & Cai, J. (2006). A semiparametric additive rates model for recurrent event data. *Lifetime Data Analysis*, **12**, 389–406.

Scheike, T. H. (2002). The additive nonparametric and semiparametric aalen model as the rate function for a counting process. *Lifetime Data Analysis*, **8**, 247–262.

Somboonsavatdee, A. & Sen, A. (2014). Parametric inference for multiple repairable systems under dependent competing risks. *Applied Stochastic Models in Business and Industry*.

Sreeja, V. N. & Sankaran, P. G. (2007). Proportional mean residual life model for gap time distributions of recurrent events. *Metron*, **LXV**, 319–336.

Strawderman, R. L. (2005). The accelerated gap times model. *Biometrika*, **92**, 647–666.

Sun, L. & Kang, F. (2013). An additive-multiplicative rates model for recurrent event data with informative terminal event. *Lifetime Data Analysis*, **19**, 117–137.

Sun, L. & Su, B. (2008). A class of accelerated means regression models for recurrent event data. *Lifetime Data Analysis*, **14**, 357–375.

Sun, L. Q., Park, D. H. & Sun, J. G. (2006). The additive hazards model for recurrent gap times. *Statistica Sinica*, **16**, 919–932.

Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer, New York.

Therneau, T. M. & Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. Springer, New York.

Tomazella, V. L. D. (2003). *Modelagem de dados de eventos recorrentes via processo de Poisson com termo de fragilidade*. Tese (Doutorado em Ciências de Computação e Matemática Computacional), Universidade de São Paulo, São Carlos.

Tsodikov, A., Ibrahim, J. & Yakovlev, A. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, **98**, 1063–1078.

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

Wang, M. C. & Chang, S. H. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association*, **94**, 146–153.

Wang, M. C., Qin, J. & Chiang, C. T. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, **96**, 1057–1065.

Wei, L. J., Lin, D. Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*, **84**, 1065–1073.

Wienke, A. (2010). *Frailty models in survival analysis*. Chapman and Hall/CRC, New York.

Xu, Y., Lam, K. F. & Cheung, Y. B. (2014). Estimation of intervention effects using recurrent event time data in the presence of event dependence and a cured fraction. *Statistics in Medicine*, **33**, 2263–2274.

Yakovlev, A. Y. & Tsodikov, A. D. (1996). *Stochastic models of tumor latency and their biostatistical applications*. World Scientific, New Jersey.

Yin, G. & Ibrahim, J. G. (2005). Cure rate models: a unified approach. *The Canadian Journal of Statistics*, **33**, 559–570.

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

Yu, B. & Peng, Y. (2008). Mixture cure models for multivariate survival data. *Computational Statistics & Data Analysis*, **52**, 1524–1532.

Zeng, D. & Lin, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, **93**, 627–640.

Zeng, D. & Lin, D. Y. (2007). Semiparametric transformation models with random effects for recurrent events. *Journal of the American Statistical Association*, **102**, 167–180.

Zeng, D. & Lin, D. Y. (2009). Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics*, **65**, 746–752.

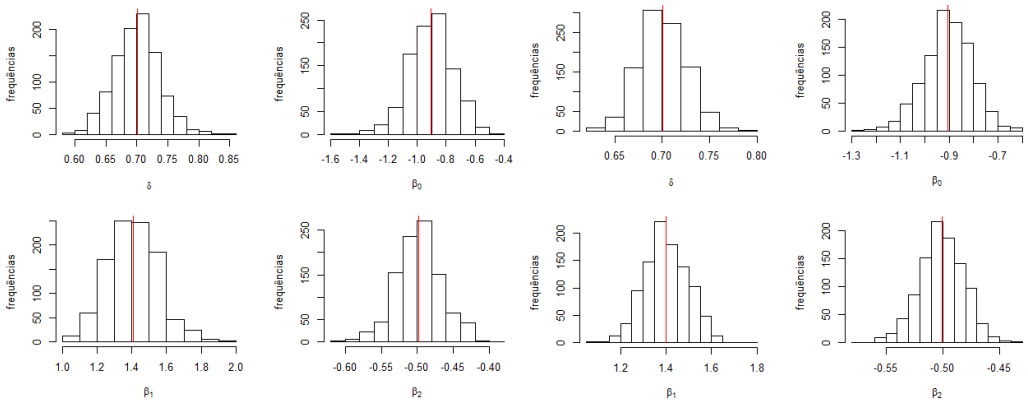
Zhao, X. & Zhou, X. (2012). Modeling gap times between recurrent events by marginal rate function. *Computational Statistics and Data Analysis*, **56**, 370–383.

Zhu, H. (2014). Non-parametric analysis of gap times for multiple event data: an overview. *International Statistical Review*, **82**, 106–122.

Apêndice A

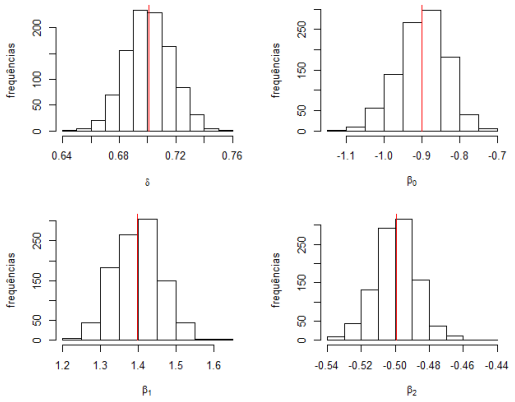
Histogramas dos parâmetros estimados para o modelo do Capítulo 3

Nesta seção apresentamos os histogramas dos parâmetros estimados com base nas 1.000 replicações de Monte Carlo para o modelo proposto no Capítulo 3, considerando as duas configurações de parâmetros (Grupo I e Grupo II) e as duas distribuições para o tempo de censura (distribuição Uniforme e Exponencial).



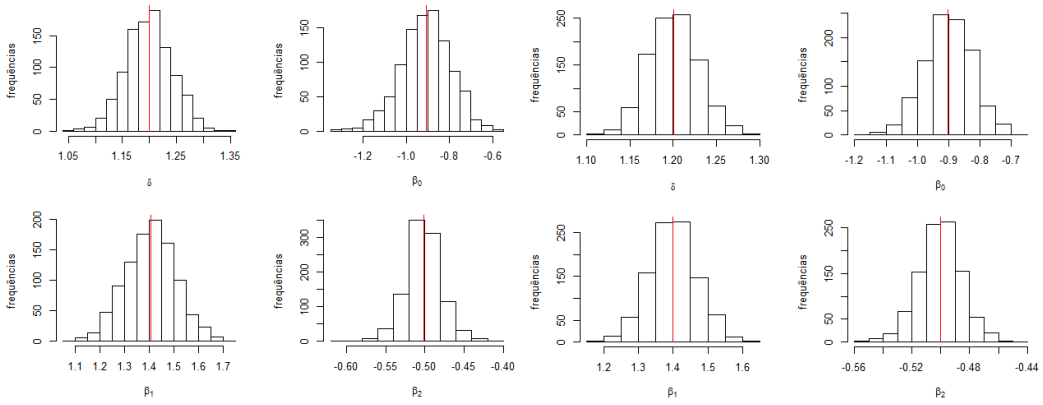
(a) n=100

(b) n=200

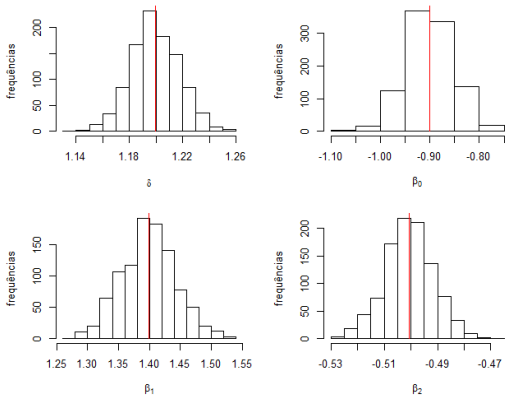


(c) n=500

Figura A.1: Histograma dos parâmetros estimados. Grupo I e tempo de censura Uniforme.

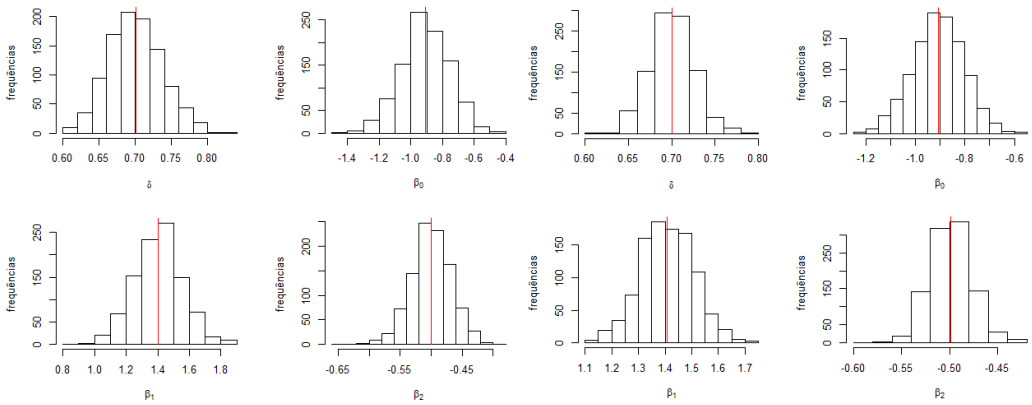


(a) n=100 (b) n=200



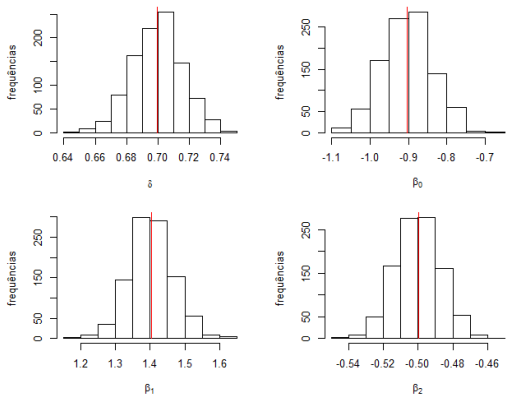
(c) n=500

Figura A.2: Histograma dos parâmetros estimados. Grupo II e tempo de censura Uniforme.



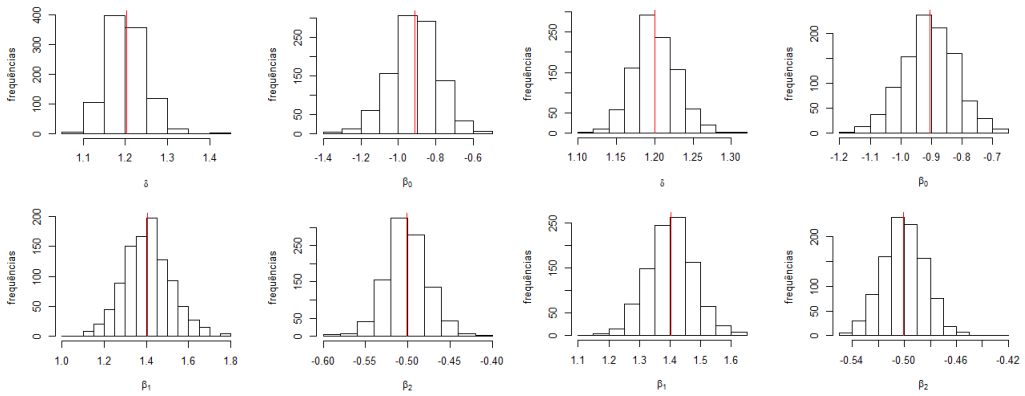
(a) n=100

(b) n=200



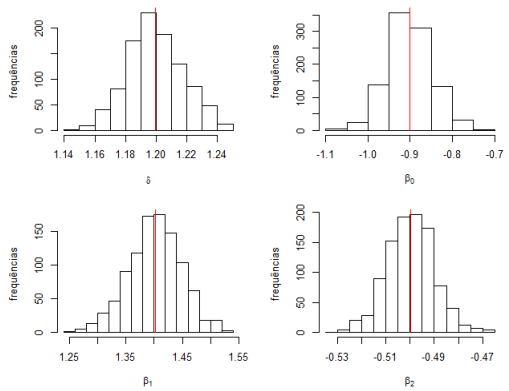
(c) n=500

Figura A.3: Histograma dos parâmetros estimados. Grupo I e tempo de censura Exponencial.



(a) n=100

(b) n=200



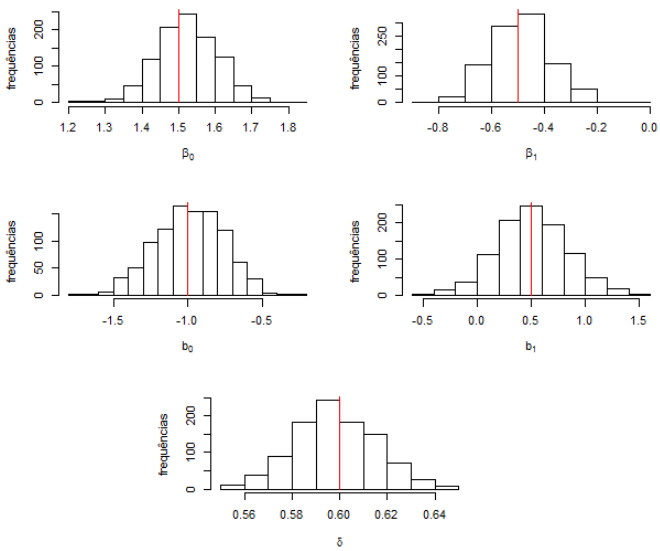
(c) n=500

Figura A.4: Histograma dos parâmetros estimados. Grupo II e tempo de censura Exponencial.

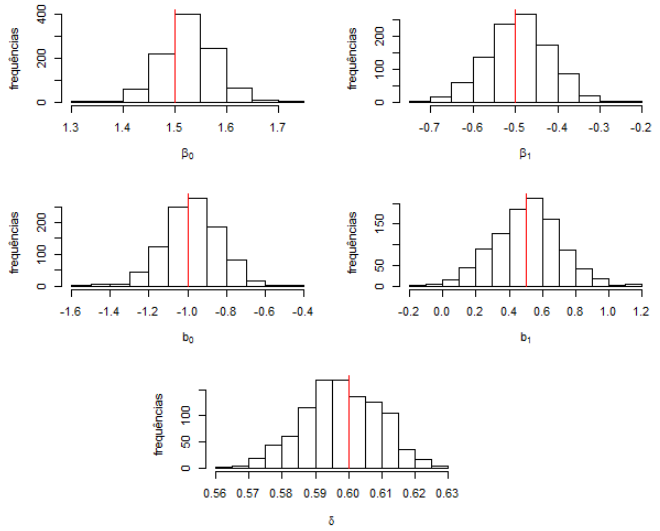
Apêndice B

Histogramas dos parâmetros estimados para o modelo do Capítulo 4

Nesta seção apresentamos os histogramas dos parâmetros estimados com base nas 1.000 replicações de Monte Carlo para o modelo com fragilidade Geométrica considerado na Seção 4.3.



(a) n=200



(b) n=500

Figura B.1: Histograma dos parâmetros estimados. Modelo com fragilidade discreta.