Universidade Federal de São Carlos
Centro de Ciências Exatas e de tecnologia
Departamento de Estatística

Estimação paramétrica do modelo de mistura padrão com fragilidade aditiva gama: Aplicação a dados de câncer de mama.

Natália Tomazella de Paula
Orientadora: Vera Tomazella

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Estimação paramétrica do modelo de mistura padrão com fragilidade aditiva gama: Aplicação a dados de câncer de mama.

Natália Tomazella de Paula

Orientadora: Profª. Drª Vera Tomazella

Trabalho de Conclusão de Curso a ser apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos

3 de Outubro de 2022

Natália Tomazella de Paula

Estimação paramétrica do modelo de mistura padrão com fragilidade aditiva gama: Aplicação a dados de câncer de mama.

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Natália Tomazella de Paula e aprovado pela banca examinadora.
São Carlos, 30 de setembro de 2022.

Banca Examinadora

- Profª. Drª Vera Tomazella
- Profº. Drº Marcio Alves Diniz
- Profº. Felipe Rodrigues

Agradecimentos

A graduação não foi um dos períodos mais fáceis da minha vida, foi um momento em que tive que encarar "sozinha"a vida adulta, mas foi um momento onde pude crescer e evoluir muito no âmbito pessoal e profissional. Sou extremamente grata por todo esse processo e por tudo que ele me proporcionou.

Agradeço infinitamente a Deus por ter me dado forças e me mantido de pé durante todos esses anos. Aos meus pais por sempre me incentivarem a dar meu melhor em tudo aquilo que faço, por terem me proporcionado um ensino maravilhoso, por toda educação que me deram, por sempre acreditarem em mim e me apoiarem em todos os momentos. E agradeço também a minha irmã que sempre esteve comigo e me ajudou nos momentos mais complicados, deixando-os sempre mais leves.

À toda minha família, principalmente aos meus avós, por todo carinho e apoio. E agradeço em especial ao tio João e a tia Vera por terem me acolhido como uma filha, por me ajudarem a ser forte para vivenciar todo esse período e por me incentivar. Tia Vera que além de ser minha tia é minha segunda mãe e orientadora. Agradeço também a minha prima Carol, que se tornou uma pessoa essencial na minha vida e me ajudou muito, principalmente nesse último semestre, com sua amizade e companheirismo.

Ao Felipe por toda ajuda que me deu para a realização desse trabalho, pela paciência e pelo conhecimento agregado.

Aos meus amigos de vida, Caroline, Bernardo e Igor, mesmo com a distância física entre a gente eles sempre se fizeram presente na minha vida, me dando apoio e torcendo pelo meu sucesso. E aos colegas de graduação, que também se tornaram amigos pra vida, Matheus Andrade, Matheus Felix, Giovanna Nesterick, Natália Paulino e Crystiane Souza, eles foram essenciais nesse período, me ajudaram em diversos momentos e eu não teria conseguido chegar até aqui sem o apoio deles.

Resumo

O câncer de mama é causado pelo crescimento desordenado de células anormais na mama, que se multiplicam de maneira rápida, agressiva e descontrolada. Essa doença afeta as mulheres do mundo todo, tanto de países desenvolvidos como subdesenvolvidos, podendo também acometer homens, sendo esta uma situação extremamente rara. Existem diversos tratamentos para sua cura, que podem variar de acordo com a fase que a doença se encontra, e, quando diagnosticada precocemente, possui uma alta taxa de cura. Modelos de taxa de cura têm sido amplamente estudados para analisar dados de tempo até o evento de interessse com a presença de uma fração de cura de pacientes. Neste contexto, o objetivo deste trabalho consiste em incorporar a fragilidade, grau de heterogeneidade induzida por fatores de risco não observáveis, no modelo de taxa de cura, para descrever este tipo de dados. O modelo proposto é uma extensão do modelo aditivo de Aalen (Aalen, 1980b) e incorpora duas parcelas da população, os indivíduos não suscetíveis e suscetíveis ao evento de interesse. Uma vantagem do modelo proposto é a possibilidade de considerar conjuntamente a heterogeneidade entre os pacientes por suas fragilidades e a presença de uma fração curada deles. Além disso serão consideradas covariáveis que influenciam a função de sobrevivência e a proporção de curados na população. Algumas propriedades destes modelos serão abordadas, bem como métodos adequados de estimação. A metodologia proposta será aplicada a um conjunto de dados de câncer de mama oriundos de estudos realizados no Hospital A.C.Camargo Cancer Center - São Paulo, Brasil.

Palavras-chave: *Análise de sobrevivência, fração de cura, modelo aditivo, fragilidade.*

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Segundo o Instituto Nacional de Câncer (INCA), o câncer de mama é uma doença que se torna cada vez mais comum entre as mulheres no Brasil e no mundo, tanto em países em desenvolvimento quanto em países desenvolvidos, respondendo por cerca de 25% dos casos novos de câncer a cada ano (Ministério da Saúde, 2016). A elevação na incidência do tumor mamário é, entre outros fatores, decorrente do acompanhamento clínico adequado, cuja importância está associada ao diagnóstico precoce do câncer. Embora a mama seja considerada uma glândula sudorípara modificada com estrutura aparentemente simples, o conhecimento de alguns componentes estruturais auxilia bastante o estudo e a classificação das diversas doenças que a acometem (Geraldo Filho, 2000). Segundo Gonçalves *et al.* (2007), existem vários fatores de risco associados ao câncer de mama e eles têm sido investigados, dado o aumento dessa doença que cada vez mais vem se tornando uma preocupação constante na vida das mulheres brasileiras.

O tratamento varia de acordo com o estadiamento da doença, suas características biológicas, bem como das condições da paciente (idade, status de menopausa e comorbidades). O prognóstico do câncer de mama depende da extensão da doença. Quando a doença é diagnosticada no início, o tratamento tem maior potencial curativo. Entretanto, quando há evidências de metástases (a doença se espalhou), o tratamento tem por objetivos principais prolongar a sobrevida e melhorar a qualidade de vida da paciente. As modalidades de tratamento do câncer de mama podem ser divididas em duas, sendo a primeira o tratamento local, onde é feito um tratamento mais simples com cirurgia e radioterapia, e a segunda o tratamento sistêmico, que é um tratamento mais árduo para a paciente e é feito quando a doença já se espalhou pelo corpo. O tratamento consiste em sessões de quimioterapia, hormonioterapia e terapia biológica (IMAMA, Porto Alegre,

2014).

Diversas pesquisas na área médica tem abordado o câncer, pode-se citar alguns trabalhos como: Farewell (1986) e Peng e Dear (2000) estudaram o câncer de mama; Laurie *et al.* (1989) fizeram estudos sobre câncer de cólon e Chen *et al.* (1999) apresentaram um estudo sobre o melanoma. Com os avanços na área médica, um número maior de pacientes passou a ser considerado "curado", ou imune à doença em estudo. Nestes estudos assume-se que os pacientes sejam suscetíveis ao evento de interesse (por exemplo morte ou recidiva da doença). Entretanto, com os avanços nos tratamentos de câncer e, por consequência na eficácia deles, os estudos conduzem a uma proporção de pacientes que não são suscetíveis ao evento de interesse esperado. A partir dos modelos tradicionais de sobrevivência não é possível estimar a fração de cura da população, ou seja, a proporção de indivíduos que são considerados curados. Assim, são necessários modelos estatísticos que incorporem tal fração e estes são denominados modelos de longa duração ou modelos de fração de cura.

O modelo proposto por Cox (1972) é um dos mais conhecidos e utilizados em análise de dados de sobrevivência. Entretanto, esse modelo supõe que os riscos são proporcionais, suposição que muitas vezes não é razoável. Para resolver essa limitação, alguns modelos foram propostos na literatura, dentre eles as chamadas extensões do modelo de Cox, como por exemplo o modelo de Cox estratificado, em que se assume que as taxas de falha são proporcionais somente em cada estrato (Colosimo e Giolo, 2006). Aalen (1980a) citou algumas limitações do modelo de Cox, com destaque para as suposições do modelo de Cox que podem não ser verificadas na prática. Outra limitação seria que mesmo quando as propriedades de proporcionalidade são satisfeitas não há garantia da adequação do modelo de Cox.

Um ponto bastante importante em análise de sobrevivência é o estudo de covariáveis, pois diversos fatores podem influenciar o tempo de sobrevivência de um indivíduo. Assim, incorporar covariáveis nos permite ter um modelo muito mais completo e repleto de informações valiosas. Por exemplo, se há interesse em estudar o tempo de vida de pacientes com uma determinada doença que estão recebendo um certo tratamento, outros fatores podem influenciar na cura do paciente e assim é possível encontrar novos meios de tratar a doença a partir de covariáveis.

Muitos autores contribuíram para a teoria dos modelos de longa duração, sendo Boag (1949) o pioneiro, em que o método de máxima verossimilhança foi utilizado para estimar

a proporção de sobreviventes em uma população de 121 mulheres com câncer de mama, experimento esse que teve a duração de 14 anos. Baseado na ideia de Boag (1949), Berkson e Gage (1952) propuseram um modelo de mistura com o objetivo de estimar a proporção de curados numa população submetida a um tratamento de câncer de estômago. Modelos mais complexos de longa duração, tais como Tsodikov *et al.* (1996), Chen *et al.* (1999) entre outros, surgiram com o objetivo de explicar melhor os efeitos biológicos envolvidos. Mais recentemente, Rodrigues *et al.* (2009) propuseram uma teoria unificada de longa duração, considerando diferentes causas competitivas. Milani *et al.* (2021) consideraram o modelo de mistura padrão para estimar a proporção de cura de mulheres que foram diagnosticadas com câncer e submetidas ao tratamento de quimioterapia neoadjuvante, que consiste em um tratamento realizado antes de um procedimento cirúrgico.

Os modelos de cura assumem implicitamente que todos os indivíduos que sofreram o evento de interesse pertencem a uma população homogênea. No entanto, existe um grau de heterogeneidade induzida por fatores de risco não observados. Nestas circunstâncias, é necessário considerar modelos que incorporam heterogeneidade não observável entre os indivíduos, como o modelo de fragilidade proposto por Vaupel *et al.* (1979). Os modelos de fragilidade são caracterizados pela inclusão de um efeito de aleatório, que é uma variável aleatória não observável tais como fatores ambientais, genéticos ou informações que por alguma razão não foram consideradas no planejamento. Uma forma de incorporar esse efeito aleatório é introduzi-lo na função de risco de forma aditiva ou multiplicativa. Estes modelos são extensões dos modelos de Aalen (1980a) e Cox (1972), respectivamente.

O fato de considerar a variável aleatória não observável introduzida na função de risco faz com que o modelo englobe duas fontes de variação para os dados. Hougaard (1991) mostrou que é vantajoso considerar as duas fontes de variabilidade. A primeira delas, que gera a heterogeneidade entre as observações, é causada por covariáveis individuais não observáveis que não foram incluídas no planejamento em estudo, por circunstancias práticas ou por serem conhecidas como sendo fatores de risco. A segunda fonte de variação é proveniente das covariáveis comuns a indivíduos de um mesmo grupo, que quando não observadas geram dependência entre os tempos de sobrevida. Uma abordagem paramétrica para modelos aditivos com fragilidade foi apresentada por Tomazella (2003) e mais adiante nesta mesma estrutura Tomazella *et al.* (2006) consideraram um procedimento de inferência Bayesiana.

1.1 Objetivo

Neste trabalho será estudado um modelo de regressão paramétrico que é uma extensão do modelo de aditivo de Aalen (1980b), denominado "Modelo de fragilidade Aditivo Gama com Proporção de Cura". Nesta proposta, supõem-se que a variável de fragilidade do modelo proposto tem distribuição Gama e o modelo de fração de cura considerado será o modelo de mistura padrão (Berkson e Gage (1952)). Nesta proposta covariáveis observadas que podem influenciar no tempo de sobrevivência serão incorporaradas na modelagem.

A metologia será analisada considerando um conjunto de dados de mulheres diagnosticadas com câncer de mama e submetidas a um tratamento de quimioterapia neo-adjuvante entre os anos de 2001 a 2013, fornecido pelo Hospital A.C.Camargo Câncer Center-São Paulo, Brasil.

1.2 Organização do Trabalho

Este trabalho está organizado da seguinte forma. No Capítulo 2, será apresentada uma revisão da literatura sobre análise de sobrevivência, abrangendo desde conceitos básicos até metodologias de modelagem de riscos aditivos e multiplicativos. No Capítulo 3, será introduzido o conceito de modelo de fragilidade. No Capítulo 4, o desenvolvimento da modelo de mistura padrão com fragilidade aditiva será apresentado e uma aplicação da metodologia proposta à dados reais. Por fim, no Capítulo 5 será apresentado conclusão do trabalho.

Capítulo 2

Revisão da Literatura

Neste capítulo será feita uma revisão dos conceitos básicos da metodologia de análise de sobrevivência, que serão essenciais para o desenvolvimento do trabalho. Todas as informações citadas a seguir foram obtidas de Colosimo e Giolo (2006).

2.1 Conceitos básicos de Análise de Sobrevivência

A análise de sobrevivência tem como principal objetivo estudar o tempo até a ocorrência de um determinado evento de interesse, conhecido como tempo de falha, que pode estar relacionado à durabilidade de um equipamento eletrônico, bem como a morte ou cura de um paciente.

Uma das principais características dos dados utilizados em sobrevivência é a presença de censura, que é a observação parcial da resposta. Ela pode ocorrer por diversos motivos, principalmente pela perda de acompanhamento do paciente e não ocorrência do evento de interesse. É importante usar esse tipo de análise em dados como esses pois, ainda que as informações não estejam completas, elas são relevantes na inferência a respeito do tempo de vida de indivíduos fora da amostra e para evitar que resultados viesados, não condizentes com a realidade da população, surjam no decorrer do estudo.

2.1.1 Censura

Conforme citado anteriormente, a censura é a principal característica desse tipo de estudo. Pode-se observar na prática três razões usuais para ocorrência de censura (ver figura 4.2) .

- O evento de interesse não ocorre para um determinado indivíduo antes do final do estudo;

- Perdeu-se o contato com o indivíduo;

- O indivíduo é retirado do estudo, por exemplo porque morreu de outra causa ou porque interrompeu o tratamento.

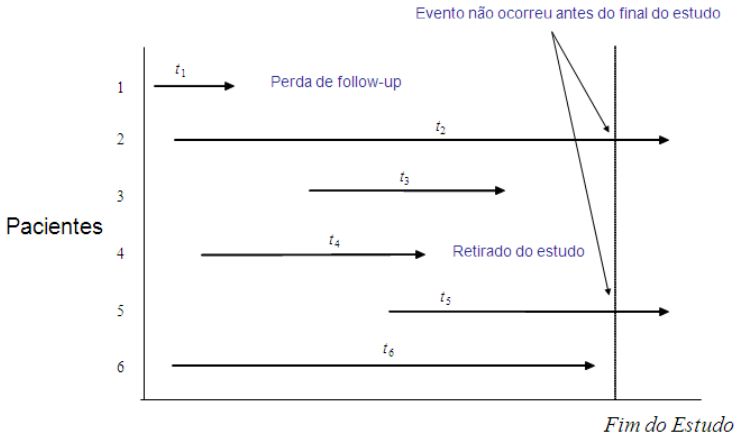

Figura 2.1: Exemplo de causas de censuras.

A seguir serão definidas as censuras mais comuns em estudos de análise de sobrevivência

Censura do Tipo I

Define-se censura do tipo I como aquela em que o término do estudo ocorre somente após um tempo pré-estabelecido, ou seja, há uma data final já estipulada. Em casos como este só é possível ter conhecimento a respeito do tempo de vida do paciente caso ele venha a óbito antes do término do estudo.

Censura do Tipo II

A censura do tipo II é caracterizada pelo estudo que é finalizado quando ocorre um número pré-estabelecido do evento de interesse. Ou seja, os indivíduos que não o apresentarem terão seu tempo censurado.

Censura Aleatória

A censura aleatória é comumente observada na área médica, visto que se trata de casos que não são possíveis de ser controlados pelo pesquisador. Ela ocorre em casos que o indivíduo é retirado do estudo sem ter apresentado a falha. Isso pode ocorrer por motivos de morte por uma razão diferente da que está sendo estudada ou porque um indíviduo deixou o estudo.

Além dos tipos de censura apresentados acima, existem definições para os mecanismos de censura.

Censura à Direita

A censura à direita ocorre quando o tempo de ocorrência do evento de interesse está à direita do tempo que foi registrado, ou seja, pode ocorrer devido a perda de acompanhamento do indivíduo ou término do estudo.

Censura à Esquerda

Quando o tempo registrado é maior que o tempo de falha, denomina-se censura à esquerda, ou seja, o indivíduo já passou pelo evento de interesse quando foi observado pela primeira vez.

Censura Intervalar

A censura intervalar acontece frequentemente, pois ocorre quando não se sabe o tempo exato em que o indíviduo passou pelo evento de interesse, tem-se conhecimento apenas do intervalo de tempo em que este ocorreu.

Os dados de sobrevivência são representados pelo par (δ_i, t_i), sendo t_i o tempo de falha, δ_i a indicadora de censura e i o $i-ésimo$ indivíduo do estudo. Sendo assim, tem-se:

$$\delta_i = \begin{cases} 1 & , \text{ se } t_i \text{ for o tempo até a falha;} \\ 0 & , \text{ se } t_i \text{ for o tempo até a censura.} \end{cases}$$

2.1.2 Funções de Interesse

A variável mais importante na análise de sobrevivência é o tempo de falha, denotado por T. Esta é uma variável aleatória não-negativa e, assumidamente, contínua. É especificada pela sua função de sobrevivência e função de taxa de falha (ou risco).

Considerando que a variável aleatória T, $T \geq 0$, tenha função de densidade de probabilidade denotada por $f(t)$, define-se esta como o limite da probabilidade de um indivíduo falhar no intervalo de tempo $[t, t + \Delta t)$ por unidade e pode ser expressa como

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

Consequentemente, sua função de distribuição acumulada é dada por:

$$F(t) = \mathbb{P}(T \leq t) = \int_0^t f(s)ds. \tag{2.2}$$

Nota-se que a Equação (2.2) retorna a probabilidade do indivíduo estar vivo antes de um determinado tempo t.

Porém, conforme dito anteriormente, o objetivo da análise de sobrevivência é estipular a probabilidade do indíviduo sobreviver após um determinado tempo. Sendo assim a função de sobrevivência será dada por

$$S(t) = \mathbb{P}(T \geq t) = \int_t^\infty f(s)ds = 1 - F(t). \tag{2.3}$$

Esta função possui as seguintes propriedades:

1. $S(t)$ é decrescente;

2. $S(0) = 1$;

3. $\lim_{t \longrightarrow \infty} S(t) = 0$.

A Figura 2.2 mostra a forma da curva de sobrevivência.

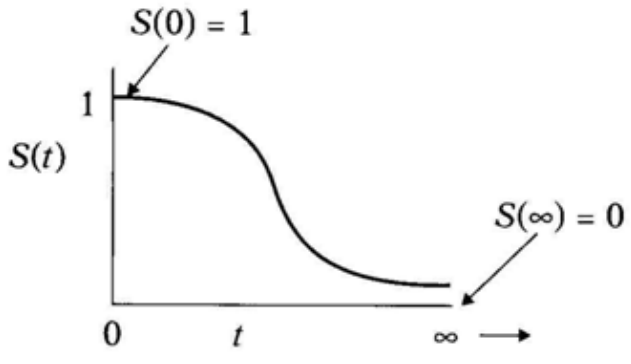

Figura 2.2: Exemplo de curva de sobrevivência.

Sabendo que o indivíduo em estudo sobreviveu até o tempo t, tem-se interesse em verificar se ele irá falhar no intervalo de tempo $[t, t + \Delta t)$, com $\Delta t \to 0$, ou seja, verificar a taxa de falha instânea, também chamada de função de risco. Esta é dada por

$$\lambda(t) = \lim_{\Delta t \longrightarrow 0} \frac{\mathbb{P}(t < T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{2.4}$$

O comportamento da função de risco pode variar de acordo com a distribuição que for atribuída à variável aleatória T. Ela pode assumir formas contantes, decrescentes, crescentes e formas não monótonas, como a "forma de banheira"ou "U"que representa, normalmente, a função de risco do tempo até a morte dos seres humanos.

Por último, tem-se a função de risco acumulado, que informa a taxa de falha acumulada do indivíduo. Ela é definida por:

$$\Lambda(t) = \int_0^t \lambda(s) ds. \tag{2.5}$$

Vale ressaltar que existem relações importantes entre as funções definidas anteriormente.

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d \log(S(t))}{dt}, \tag{2.6}$$

$$\Lambda(t) = -\log(S(t)) \tag{2.7}$$

e

$$S(t) = \exp(-\Lambda(t)). \tag{2.8}$$

Além dessas relações, existem outras que podem ser construídas a partir das funções encontradas em Colosimo e Giolo (2006).

2.1.3 Estimador de Kaplan-Meier

Nos estudos de análise de sobrevivência são apresentadas classes de estimadores para a função de sobrevivência com a presença de censura. Existe a classe de estimadores não paramétricos, que não atribuem à variável de interesse, tempo até a ocorrência do evento,

um modelo estatístico paramétrico. Alguns deles são: o estimador de Kaplan-Meier, o estimador de Nelson-Aalen, proposto por Nelson (1972), e suas propriedades estudas por Aalen (1978). Outro estimador também bastante conhecido é a tabela de vida ou atuarial, uma das técnicas mais antigas.

O estimador de Kaplan-Meier, conhecido por estimador limite-produto, é uma adaptação da função de sobrevivência empírica, pois leva em consideração a presença de censura. Ele é calculado pela seguinte expressão:

$$\widehat{S}(t) = \prod_{j:\ t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:\ t_j<t} \left(1 - \frac{d_j}{n_j} \right), \tag{2.9}$$

em que $t_1, t_2, ..., t_r$ são considerados os k tempos de falhas ordenados de forma crescente, d_j é o número de falhas no tempo t_j, $j = 1, ..., k$, e n_j é o número de indivíduos expostos ao risco em t_j, $j = 1, ..., k$.

De acordo com Colosimo e Giolo (2006), as principais propriedades desse estimador são:

1. é não-viesado para amostras grandes;

2. é fracamente consistente;

3. converge assintoticamente para um processo gaussiano;

4. é estimador de máxima verossimilhança de $S(t)$.

O estimador de Nelson-Aalen (Nelson, 1972) é mais recente que o anterior e se baseia na seguinte função:

$$S(t) = \exp\{-\Lambda(t)\} \tag{2.10}$$

em que $\Lambda(t)$ é a função de risco acumulado.

Esse estimador provou propriedades assintóticas usando processos de contagem e possui a seguinte forma:

$$\hat{\Lambda}(t) = \sum_{j:t_j<t} \left(\frac{d_j}{n_j} \right), \tag{2.11}$$

em que d_j e n_j possuem a mesma definição já mostrada anteriormente.

A curva de sobrevivência do estimador Kaplan-Meier tem forma de degrau com mudança no valor da função de sobrevivência para cada valor de falha observado. A Figura 2.3 descreve uma forma típica da curva de Kaplan-Meier.

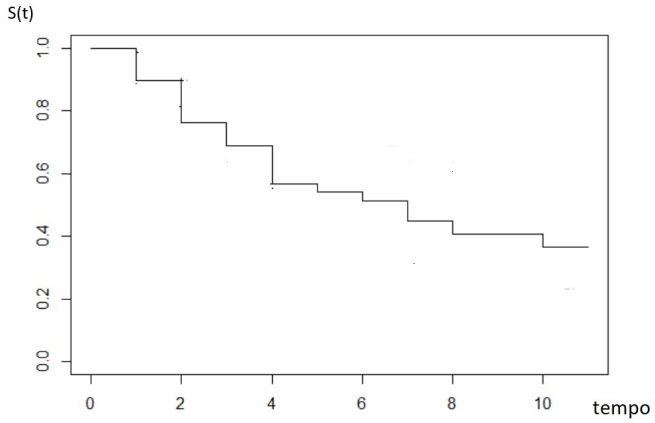

Figura 2.3: Função Sobrevivência de Kaplan-Meier.

2.2 Modelos Paramétricos

Para os métodos de estimação paramétrica é necessário que se assuma uma distribuição de probabilidade que descreva adequadamente os dados de sobrevivência. Existem diversas distribuições que são consideradas apropriadas para descrever a variável de interesse, tempo até a falha. A escolha dessa é de extrema importância, visto que cada distribuição pode gerar estimadores diferentes e seu uso inadequado pode gerar grandes erros.

Tendo isso em vista, neste seção serão apresentados alguns dos principais modelos utilizados na área: Exponencial, Weibull e Gama.

2.2.1 Distribuição Exponencial

A distribuição exponencial é conhecida por ser um dos modelos probabilísticos mais simples para descrever o evento de interesse da área de sobrevivência e por ser a única distribuição contínua que "não tem memória", ou seja, o tempo de vida não afeta na sobrevivência, o que faz com que sua função de risco seja constante.

Essa distribuição possui apenas um parâmetro, θ, e sua função de densidade de pro-

babilidade, considerando T a variável aleatória tempo de falha, é dada por:

$$f_E(t) = \theta \exp\{-\theta t\}, \ t \geq 0 \text{ e } \theta > 0. \quad (2.12)$$

A função de sobrevivência e risco são dadas, respectivamente, por:

$$S_E(t) = \exp\{-\theta t\}, \ t \geq 0 \text{ e } \theta > 0 \quad (2.13)$$

e

$$\lambda_E(t) = \theta, \ t \geq 0 \text{ e } \theta > 0. \quad (2.14)$$

A Figura 2.4 mostra o comportamento da função de densidade de probabilidade, de sobrevivência e de risco, respectivamente, para diferentes valores do parâmetro θ.

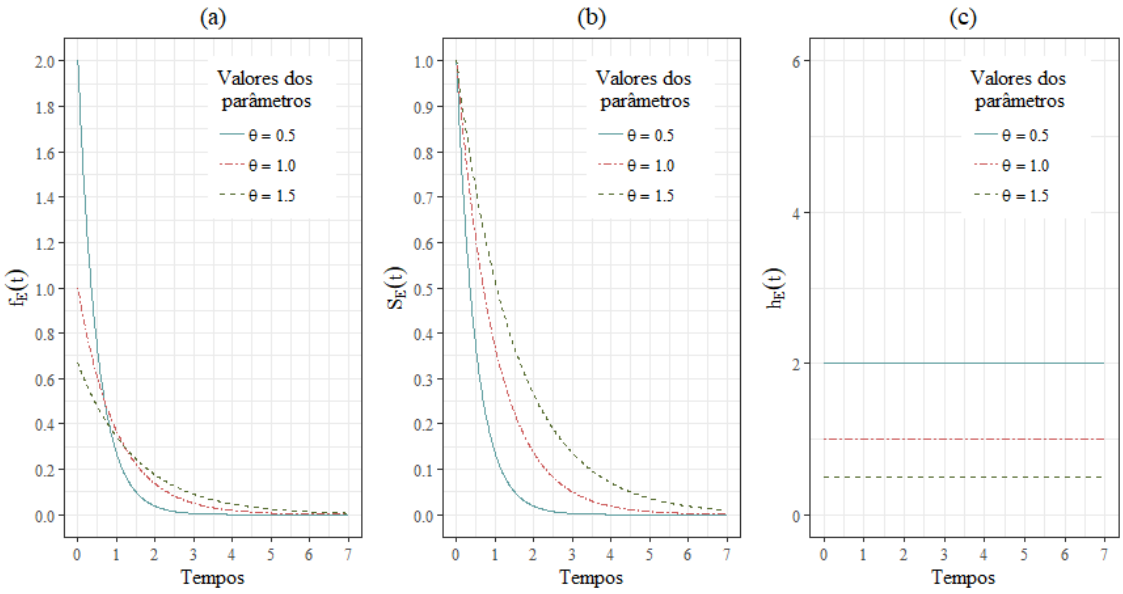

Figura 2.4: Funções densidades (a), sobrevivência (b) e taxa de falha (c) para a distribuição Exponencial.

2.2.2 Distribuição Weibull

A distribuição Weibull foi proposta por Weibull (1939) e sua teve aplicabilidade estudada pelo mesmo autor, Weibull (1951). É bastante conhecida por apresentar uma grande variedade de formas, em que todas elas possuem uma propriedade: a função de taxa de falha é sempre monótona, ou seja, ela pode ser crescente, decrescente ou constante.

Essa distribuição possui dois parâmetros, γ, parâmetro de forma, e θ, parâmetro de escala. A função de densidade de probabilidade para essa, quando se considera T a variável

aleatória tempo de falha, é dada por:

$$f_W(t) = \gamma\theta t^{\gamma-1} exp[-\theta t^{\gamma}],\ t \geq 0 \text{ e } \theta > 0,\ \gamma > 0. \tag{2.15}$$

A função de distribuição acumulada da Weibull é dada por:

$$F_W(t) = 1 - exp[-\theta t^{\gamma}]. \tag{2.16}$$

A função de sobrevivência para esse modelo é dada por:

$$S_W(t) = 1 - F(t) = exp[-\theta t^{\gamma}]. \tag{2.17}$$

Utilizando as relações existentes entre essas funções, pode-se obter a função de risco

$$\lambda_W(t) = \frac{f(t)}{S(t)} = \frac{\gamma\theta t^{\gamma-1} exp[-\theta t^{\gamma}]}{exp[-\theta t^{\gamma}]} = \gamma\theta t^{\gamma-1}. \tag{2.18}$$

Quando o parâmetro de forma, γ, é alterado, obtém-se uma grande variedade de comportamentos da distribuição Weibull, podendo mesmo até chegar em outras distribuições, que serão conhecidas como casos particulares dessa, como mostrado abaixo.

- $\gamma = 1$: a Weibull é uma distribuição Exponencial;

- $\gamma = 2$: a Weibull é uma distribuição Rayleigh;

- $\gamma = 2,5$: a Weibull aproxima-se da distribuição Log-Normal;

- $\gamma = 3,6$: a Weibull aproxima-se da distribuição Normal.

A Figura 2.5 mostra que a forma das funções definidas acima está sucetível a mudanças de acordo com o valor dos parâmetros.

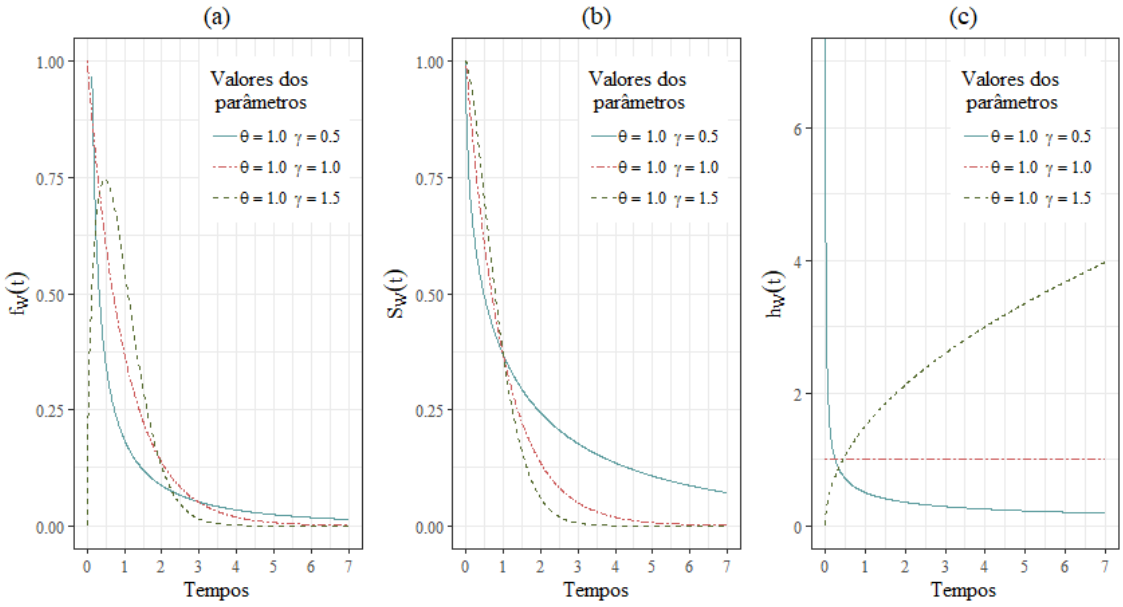

Figura 2.5: Funções densidades (a), sobrevivência (b) e taxa de falha (c) para a distribuição Weibull.

2.2.3 Distribuição Gama

A distribuição Gama é principalmente utilizada em problemas de confiabilidade devido ao fato de se ajustar de forma adequada às diversas situações que a área estuda. Quando se assume a presença de efeitos aleatórios, como nos modelos de fragilidade, é utilizada frequentemente para modelar esses componentes (Colosimo e Giolo, 2006).

Essa distribuição possui dois parâmetros, um parâmetro de forma, e um parâmetro de escala. A função de densidade de probabilidade quando se considera T a variável aleatória tempo de falha e os dois parâmetros iguais $T \sim gama(\alpha, \alpha)$, é dada por:

$$f_G(t) = \frac{\alpha^\alpha}{\Gamma(\alpha)} t^{\alpha-1} exp\{-\alpha t\}, \tag{2.19}$$

em que $t > 0$ e $\alpha > 0$.Neste caso $E(T) = 1$ e $V(T) = 1/\alpha$.

A respectiva função de sobrevivência é dada por:

$$S_G(t) = \int_t^\infty \frac{\alpha^\alpha}{\Gamma(\alpha)} u^{\alpha-1} exp\{-\alpha u\} du. \tag{2.20}$$

A função de taxa de falha pode ser obtida por meio da seguinte relação:

$$\lambda_G(t) = \frac{f(t)}{S(t)}. \tag{2.21}$$

A Figura 2.6 mostra as diferentes formas que as funções definidas acima podem assumir. Elas foram representadas graficamente, levando em consideração diferentes valores dos parâmetros.

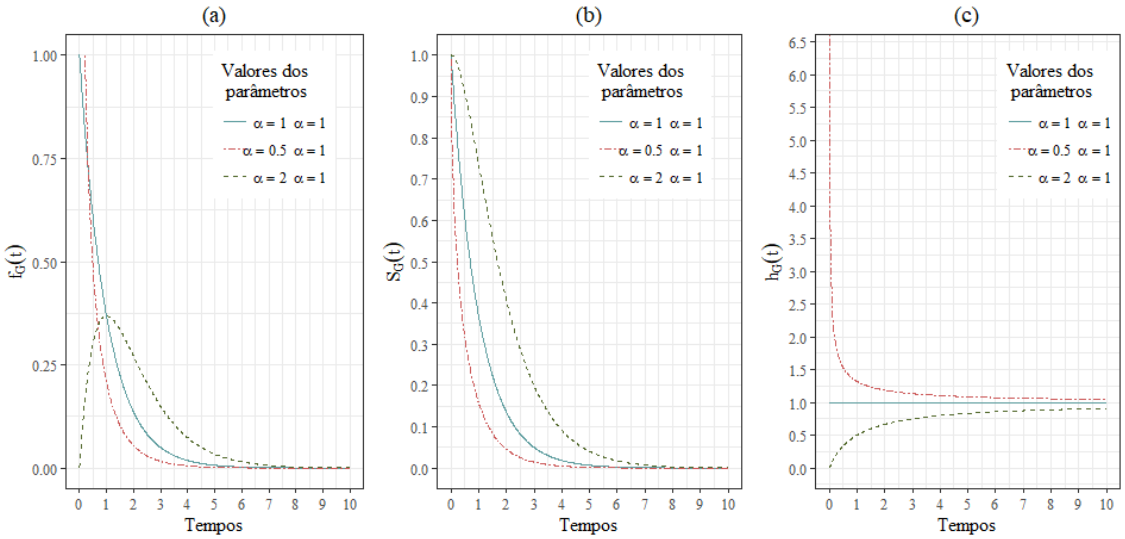

Figura 2.6: Funções densidades (a), sobrevivência (b) e taxa de falha (c) para a distribuição Gama.

2.2.4 Estimação por Máxima Verossimilhança

Existem situações em análise de sobrevivência em que existe o interesse de estimar a função de sobrevivência supondo que o tempo de falha segue uma distribuição de probabilidade.

O método da máxima verossimilhança é capaz de incorporar censuras e possui ótimas propriedades para amostras grandes e por isso é um método muito utilizado em análise de sobrevivência.

Para dados sem censura, a função de verossimilhança é definida por:

$$L(\pi) = \prod_{i=1}^{n} f(t_i; \pi), \tag{2.22}$$

em que T é a variável aleatória representando o tempo de vida com função densidade de probabilidade $f(t;\pi)$ e π é o vetor de parâmetros do modelo escolhido.

Porém, em análise de sobrevivência são utilizados dados com presença de censura e estes trazem informações importantes para o modelo, pois quando há presença de censura

sabe-se que o tempo de falha do indivíduo é maior do que aquele em que foi censurado. Sendo assim, sua contribuição para $L(\pi)$ é dada pela função de sobrevivência $S(t)$.

$$L(\pi) = \prod_{i=1}^{n} [f(t_i;\pi)]^{\delta_i} [S(t_i;\pi)]^{1-\delta_i} = \prod_{i=1}^{n} [\lambda(t_i;\pi)]^{\delta_i} S(t_i;\pi), \tag{2.23}$$

em que δ_i é a variável indicadora de falha. A expressão acima é válida para as censuras do tipo I, II, aleatória e também sob a suposição que o mecanismo de censura é não informativa.

Os estimadores de máxima verossimilhança são os valores de π que maximizam $L(\pi)$, ou de forma equivalente, $\ell(\pi) = \log[L(\pi)]$. Em muitos modelos, os estimadores podem ser encontrados resolvendo-se o sistema de equações

$$U(\hat{\pi}) = \frac{\partial \ell(\hat{\pi})}{\partial \pi} = 0. \tag{2.24}$$

Normalmente, o estimador de máxima verossimilhança não possui uma expressão fechada por sua complexidade. Então, faz-se necessário o uso de métodos numéricos para fazer a estimação.

2.3 Modelo de Riscos Proporcionais de Cox

Os modelos de regressão paramétricos exigem uma suposição de uma distribuição estatística para o tempo de sobrevivência. Contudo, se a suposição não for adequada, as estimativas podem ser pouco confiáveis. Sendo assim, com o objetivo de encontrar um modelo mais flexível, Cox (1972) propôs um modelo, denominado Modelo de Risco Proporcional de Cox.

Nesse modelo, assume-se que os tempos $t_i = 1, ..., n$ são independentes e que o risco individual é dado por,

$$\lambda(t|x_i) = \lambda_0(t) g(\mathbf{x_i'}\boldsymbol{\beta}) \tag{2.25}$$

em que $\lambda_0(t)$ é a função de risco básico, ou seja, é o risco de um indivíduo com o vetor de covariáveis nulo, x é o vetor de covariáveis e β é o vetor de parâmetros associados às covariáveis x. Já o componente paramétrico da função $(g(\mathbf{x_i'}\boldsymbol{\beta}))$ é sempre não-negativo e

frequentemente utilizado na seguinte forma multiplicativa, sendo,

$$g(\mathbf{x_i'}\boldsymbol{\beta}) = \exp\{\mathbf{x_i'}\boldsymbol{\beta}\} = \exp\{\mathbf{x_1}\boldsymbol{\beta_1} + \mathbf{x_2}\boldsymbol{\beta_2} + ... + \mathbf{x_k}\boldsymbol{\beta_k}\}, \tag{2.26}$$

onde $\boldsymbol{\beta}$ é o vetor de parâmetros associados às covariáveis $\mathbf{x}$.

O modelo de Cox é dito semi-paramétrico (2.25), pois a função de risco de base é não-paramétrica. Sendo assim, esse modelo se torna mais flexível do que o modelo paramétrico devido à presença da função de base.

Ademais, o modelo também é denominado de modelo de riscos proporcionais pois a razão da taxa de falha de dois indivíduos diferentes, i e j, é constante no tempo. Isto é, a razão das funções de taxa de falha dos indivíduos i e j dada por

$$\frac{\lambda_i(t|\mathbf{x})}{\lambda_j(t|\mathbf{x})} = \frac{\lambda_0(t)\exp(\mathbf{x_i'}\boldsymbol{\beta})}{\lambda_0(t)\exp(\mathbf{x_i'}\boldsymbol{\beta})} = \frac{\exp(\mathbf{x_i'}\boldsymbol{\beta})}{\exp(\mathbf{x_i'}\boldsymbol{\beta})} = \exp[\boldsymbol{\beta}(\mathbf{x_i} - \mathbf{x_j})] \tag{2.27}$$

não depende do tempo. Nesse sentido, se um indivíduo no início do estudo possui um risco de morte igual a duas vezes o risco de um segundo indivíduo, então, a razão de riscos será a mesma para todo o período de acompanhamento.

Para fazer uso do modelo de regressão de Cox, é necessário seguir uma suposição básica de que as taxas de falhas devem ser proporcionais ou, de forma equivalente para este modelo, que as taxas de falha acumuladas também sejam proporcionais.

2.3.1 Funções relacionadas a $\lambda_0(t)$

As funções relacionadas a $\lambda_0(t)$ referem-se basicamente a:

- Função de risco básico acumulada

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du. \tag{2.28}$$

- Função de sobrevivência de base

$$S_0(t) = \exp\{-\Lambda_0(t)\}. \tag{2.29}$$

Dessa forma, tem-se:

- Função de risco acumulada

$$\Lambda(t|x) = \Lambda_0(t)\exp\{\mathbf{x_i'}\boldsymbol{\beta}\} \tag{2.30}$$

- Função de sobrevivência

$$S(t|x) = [S_0(t)]^{\exp\{\mathbf{x_i'}\boldsymbol{\beta}\}} \tag{2.31}$$

2.3.2 Estimação do Modelo de Cox

Os coeficientes de regressão $\beta's$ são as quantidades de maior interesse na modelagem estatística de dados, pois medem os efeitos das covariáveis sobre a função da taxa de falha. Essas quantidades devem ser estimadas a partir das observações amostrais.

Nesse contexto, é necessário um método de estimação para fazer inferências neste modelo, e o método de máxima verossimilhança é o mais utilizado nessas ocasiões. Contudo, com a presença do componente não-paramétrico $\lambda_0(t)$ na função de verossimilhança (2.23), o método torna-se inapropriado para o uso.

Logo, para solucionar o problema Cox (1975) propôs uma solução razoável que consiste em condicionar a construção da função de verossimilhança ao conhecimento da história passada de falhas e censuras para eliminar esta pertubação da verossimilhança. Esse novo método foi denominado como método de máxima verossimilhança parcial.

2.3.3 Método de Verossimilhança Parcial

Considere uma amostra de n indivíduos e que existam $k \leq n$ falhas distintas nos tempos $t_1 < t_2 < ... < t_k$. A função de verossimilhança é dada por,

$$L_P(\beta) = \prod_{i=1}^{n}\left(\frac{\exp\{\mathbf{x_i'}\boldsymbol{\beta}\}}{\sum_{j\in S(t_i)}\exp\{\mathbf{x_i'}\boldsymbol{\beta}\}}\right)^{\delta_i}, \tag{2.32}$$

onde δ_i é o indicador de censura e $S(t_i)$ é o conjunto de índices das observações sob risco no tempo t_i. Observe que condicional à história de falhas e censuras até o tempo t_i, o componente não-paramétrico $\lambda_0(t)$ desaparece.

Então, a função de verossimilhança (2.32) a ser utilizada para fazer inferências no modelo de Cox, é formada pelo produto de todos os termos associados aos tempos distintos de falha. Os valores de $\boldsymbol{\beta}$ que maximizam a função de verossimilhança parcial, $L_P(\boldsymbol{\beta})$ são

obtidos resolvendo o sistema de equações definido por $U_P(\boldsymbol{\beta}) = 0$, em que $U_P(\boldsymbol{\beta})$ é vetor de derivadas de primeira ordem da função $\log(L_P(\boldsymbol{\beta}))$, isto é,

$$U_P(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{j \in S(t_i)} x_j \exp\{\mathbf{x_i'}\hat{\boldsymbol{\beta}}\}}{\sum_{j \in S(t_i)} \exp\{\mathbf{x_i'}\hat{\boldsymbol{\beta}}\}} \right] = 0. \tag{2.33}$$

2.3.4 Estimativa das funções relacionadas a $\Lambda_0(t)$

Um estimador simples, proposto para a função de risco de base acumulada $\Lambda_0(t)$ é o estimador de Breslow, que é uma função escada com saltos nos distintos tempos de falha e é expresso por,

$$\hat{\Lambda}_0(t) = \sum_{j:t_j<t} \frac{d_j}{\sum_{I \in R_j} \exp\{\mathbf{x_i'}\hat{\boldsymbol{\beta}}\}}, \tag{2.34}$$

em que d_j é o número de falhas em t_j e R_j é o conjunto dos índices das observações sob risco no tempo j.

Por consequência, as funções de sobrevivência $S_0(t)$ e $S(t)$ podem ser estimadas por, respectivamente,

$$\hat{S}_0(t) = \exp\{-\hat{\Lambda}_0(t)\} \tag{2.35}$$

$$\hat{S}(t|x) = [\hat{S}_0(t)]^{\exp\{\mathbf{x_i'}\hat{\boldsymbol{\beta}}\}} \tag{2.36}$$

2.3.5 Interpretação dos Coeficientes

Os coeficientes $\boldsymbol{\beta}$ no modelo de regressão de Cox, medem os efeitos das covariáveis na função de risco, sendo que uma covariável pode acelerar ou desacelerar a função de risco.

Nesse sentido, quando o coeficiente $\boldsymbol{\beta}$ possui valores positivos, têm-se variáveis que contribuem para o aumento do risco, mas quando o coeficiente possui valores negativos, têm-se variáveis que contribuem para a redução do risco. Sendo assim, quando a exponencial do coeficiente, $e^{\boldsymbol{\beta}}$, possui risco relativo maiores do que 1, tem-se um sobrerisco e quando possui risco relativo entre 0 e 1, tem-se uma proteção.

2.3.6 Modelos de Cox Paramétricos

Para os modelos de riscos proporcionais paramétricos assume-se que a função de risco do indivíduo i é dada por,

$$\lambda(t|x_i) = \lambda_0(t) g(\mathbf{x_i'}\boldsymbol{\beta}), \tag{2.37}$$

e como nesse modelo λ_0 é assumida uma forma paramétrica, tem-se então diferentes formas paramétricas para o tempo de vida T, sendo a distribuição Exponencial, Weibull, Gama, Gama Generalizada e entre outros.

Nesse modelo a função de sobrevivência condicional para o indivíduo i é dada por,

$$S(t|x_i) = [S_0(t)]^{exp\{\mathbf{x_i'}\boldsymbol{\beta}\}}. \tag{2.38}$$

e a função de verossimilhança é dada por,

$$L(\boldsymbol{\beta}, x) = \prod_{i=1}^{n} [\lambda_0(t) \exp\{\mathbf{x}'\boldsymbol{\beta}\}]^{\delta_i} [S_0(t_i)]^{\exp\{\mathbf{x}'\boldsymbol{\beta}\}} \tag{2.39}$$

em que $\lambda_0(t)$ é a função de risco de base e $S_0(t)$ é a função de sobrevivência de base.

Considerando que a função de risco base tem distribuição expoenencial tem-se então o **modelo de riscos proporcionais exponencial** com função de risco de base dada por:

$$\lambda_0(t) = \theta \tag{2.40}$$

e com isso, a função de risco e confiabilidade condicionais do modelo de Cox são dadas por,

$$\lambda(t|x_i) = \theta \exp(\mathbf{x_i'}\boldsymbol{\beta}) \tag{2.41}$$

$$S(t|x_i) = [\exp(-\theta t)]^{\exp(\mathbf{x_i'}\boldsymbol{\beta})}. \tag{2.42}$$

Nesse contexto, a função de verossimilhança dos modelos de riscos proporcionais exponencial, será de,

$$L(\boldsymbol{\beta}, x) = \prod_{i=1}^{n} [\theta \exp(\mathbf{x}'\boldsymbol{\beta})]^{\delta_i} [\exp(-\theta t)]^{\exp(\mathbf{x}'\boldsymbol{\beta})} \tag{2.43}$$

onde δ_i é o indicador de falha.

Já nos **modelos de riscos proporcionais Weibull**, têm-se a função de risco base de Weibull que é dada por,

$$\lambda_0(t) = \gamma\theta t^{\gamma-1} \tag{2.44}$$

e assim, a função de risco e confiabilidade do modelo de Cox são dadas por,

$$\lambda(t|x_i) = \gamma\theta t^{\gamma-1}\exp(\mathbf{x_i'}\boldsymbol{\beta}) \tag{2.45}$$

$$S(t|x_i) = \left[\exp\left(-\theta t^{\gamma}\right)\right]^{\exp(\mathbf{x_i'}\boldsymbol{\beta})}. \tag{2.46}$$

A função de verossimilhança dos modelos de riscos proporcionais Weibull é dado por,

$$L(\boldsymbol{\beta}, x) = \prod_{i=1}^{n}\left[\gamma\theta t^{\gamma-1}\exp(\mathbf{x'}\boldsymbol{\beta})\right]^{\delta_i}\left[\exp\left(-\theta t^{\gamma}\right)\right]^{\exp(\mathbf{x'}\boldsymbol{\beta})}, \tag{2.47}$$

onde δ_i é o indicador de falha.

2.4 Modelo de risco aditivo de Aalen

Considerando situações em que não se impõe que as funções de risco sejam proporcionais, tem-se a segunda classe de modelos de risco em análise de sobrevivência, conhecida por modelos de riscos aditivos. O modelo proposto por Aalen (1980a) é aquele em que o efeito das covariáveis é expresso aditivamente na função de risco e é definido por:

$$\lambda\left(t|\mathbf{x}_i\right) = \lambda_0\left(t_i\right) + G\left(\mathbf{x}_i, \boldsymbol{\beta}\right), \tag{2.48}$$

em que $\lambda_0(\cdot)$ é a função de risco de base, e $G(\cdot)$ é uma função positiva e $\boldsymbol{\beta}$ é um vetor p dimensional de parâmetro desconhecidos de efeitos associados a $\mathbf{x}_i$.

Assumindo que a função $g(\cdot) = \mathbf{x_i'}\boldsymbol{\beta}$ tem-se que (2.48) pode ser escrita como:

$$\lambda(t|\mathbf{x}_i) = \lambda_0(t) + \mathbf{x_i'}\boldsymbol{\beta}. \tag{2.49}$$

Note que essa modelagem assumindo a função $g(\cdot)$ como função identidade pode apresentar valor negativo para a função de risco, o que é uma desvantagem em relação à abordagem multiplicativa que assume uma função exponencial para $g(\cdot)$ multiplicando o risco de base. Por se tratar de uma abordagem paramétrica deve-se supor uma distribuição de probabilidade para o risco de base.

A forma mais geral da função de risco para o modelo aditivo (2.48) é dada da seguinte

forma:

$$\lambda(t|\mathbf{x}_i(t)) = \boldsymbol{\beta_0}(t) + \sum_{j=1}^{p} \boldsymbol{\beta_j}(t) x_{ij}(t),$$

em que $x_{ij}(t)$ são as covariáveis que podem depender do tempo, com $i = 1, \ldots, n$ e $j = 1, \ldots, p$, onde n é o número de indivíduos no estudo e p é o número de covariáveis no modelo. Os $\boldsymbol{\beta_j}(t)$ são funções do tempo desconhecidas que medem a influência das respectivas covariáveis e $\boldsymbol{\beta_0}(t)$ pode ser visto como uma função de risco base ($\lambda_0(t)$).

O modelo aditivo de Aalen apresenta algumas caracteríscas importantes, como o fato do efeito das covariáveis ser aditivo na função de risco e possibilitar o monitoramento de covariáveis dependentes do tempo. Essa modelagem também permite que os parâmetros e as covariáveis variem com o tempo e nenhuma forma paramétrica particular é assumida para os $\boldsymbol{\beta_j(t)}$, chamados de coeficientes de risco ou funções de regressão. E, por fim, por se tratar de um modelo que ajusta funções e não parâmetros, é considerado um modelo não-paramétrico e flexível. A desvantagem de utilizar essa metodologia é que ela estima tanto valores negativos como positivos para a função de risco.

A estimação desse modelo não é trivial, normalmente, utiliza-se como uma alternativa a estimação da função de regressão acumulada, dada por:

$$B_j(t) = \int_0^t \boldsymbol{\beta_j}(u) \mathrm{d}u,$$

e o estimador proposto por Aalen para estimar $B_j(t)$, é dado por:

$$\hat{B}(t) = \sum_{t_i < t} Z(t_i) I(t_i),$$

em que, $t_1 < t_2 < \cdots < t_k$ são os tempos de falha ordenados, $\hat{B}(t)$ é o vetor dos $\hat{B}_j(t)$, $I(t_i)$ é um vetor com o i-ésimo elemento igual a 1 se o evento ocorre para o indivíduo i no tempo t e é zero caso contrário e $Z(t_i)$ é a inversa generalizada de $X(t_i)$ que é uma matriz $n \times (p+1)$ composta por vetores do tipo $\mathbf{x}_i(t) = (1, x_{i1}(t), \ldots, x_{ip}(t))'$ na i-ésima linha se a falha não ocorrreu para o indivíduo i e se ele ainda estiver no estudo, e contém apenas zeros na linha i se o i-ésimo indivíduo não estiver em risco no tempo t.

Os $\hat{B}(t)$ podem ser vistos como funções empíricas que descrevem a influência da j-ésima covariável, e a inclinação da função de regressão acumulada $\hat{B}_j(t)$ indica como a covariável influencia a função de risco ao longo do tempo. Ou seja, uma inclinação positiva em relação ao eixo do tempo indica que naquele período o aumento dos valores

da covariável levou ao aumento do risco, e se a inclinação for negativa significa que naquele período o aumento dos valores da covariável levou à redução do risco, e, por fim, se não houver inclinação (comportamento constante) quer dizer que a covariável relacionada tem efeito constante no risco.

O estimador da função de risco acumulada, quando todas as covariáveis já foram fixadas no tempo inicial do estudo, é dado por:

$$\hat{\Lambda}(t|\mathbf{x}) = \hat{B}_0(t) + \sum_{j=1}^{p} \hat{B}_j(t)x_j,$$

sendo os valores de t tais que a matriz $X(t)$ é não-singular. A função de sobrevivência estimada é dada pela seguinte relação:

$$\hat{S}(t|\mathbf{x}) = \exp\{-\hat{\Lambda}(t|\mathbf{x})\}.$$

Caso não seja utilizada a abordagem paramétrica mostrada, o modelo de risco aditivo proposto por Lin e Ying (1994) tem uma abordagem semi-paramétrica, assumindo para função de risco de base uma função não-negativa e considerando as funções de regressão $\boldsymbol{\beta_j}(t)$ constantes ao longo do tempo, dessa forma tem-se o modelo de Aalen dado em (2.49). Para esse modelo a interpretação dos parâmetros é mais direta do que no modelo original de Aalen, entretanto ainda mantém a desvantagem de admitir a estimação da função de risco com valores negativos.

2.5 Considerações finais

Este capítulo teve o intuito de apresentar os conceitos básicos da análise de sobrevivência, como suas principais funções, definição de censura e os principais modelos estatísticos paramétricos. Além disso, foram abordados os conceitos de modelo de riscos proporcionais de Cox e modelo de risco aditivo de Aalen, conceitos importantes para a contrução do modelagem de modelo de mistura padrão com fragilidade aditiva. No próximo capítulo será abordado o tema a respeito do modelo de fragilidade

Capítulo 3

Modelo de Fragilidade

O modelo de fragilidade é caracterizado pela utilização de um efeito aleatório, ou seja, de uma variável aleatória não observável, que representa as informações que não podem ou não foram observadas tais como: fatores ambientais, genéticos, informações que, por algum motivo, não foram consideradas no planejamento amostral.

O modelo de fragilidade engloba duas fontes de variações: a que gera a heterogeneidade entre as observações causada por covariáveis individuais não observadas que não foram incluídas no planejamento do estudo por circunstâncias práticas, ou por não serem conhecidas como sendo fatores de risco, e aquela proveniente das covariáveis comuns a indivíduos de um mesmo grupo ou família que, quando não são observadas, geram dependência entre os tempos de eventos (Tomazella, 2003).

As fragilidades podem ser introduzidas de forma multiplicativa ou de forma aditiva na função de risco, visando assim responder as diferentes formas de avaliar a influência da heterogeneidade entre as unidades na função de risco, ou intensidade nos processos de contagem. Existem duas categorias de modelos de fragilidade, que são os modelos de fragilidades para dados univariados e multivariados.

Os modelos de fragilidade para dados de sobrevivência univariados leva em conta que a população é não homogênea. A heterogeneidade pode ser explicada por covariáveis, mas quando importantes covariáveis não são incorporadas no modelo, há uma certa heterogeneidade, embora não observada. Já os modelos de fragilidade para dados de sobrevivência multivariado é muito comum para indivíduos com eventos repetidos ou dados de eventos recorrentes ou causas competitivas (Colosimo e Giolo, 2006).

O modelo de fragilidade compartilhado remonta a Clayton (1978) e leva em consideração a associação entre os tempos de sobrevivência dos indivíduos dentro de cada

grupo. A fragilidade representa, nessses casos, um efeito aleatório que descreve o risco comum, isto é, a fragilidade compartilhada por indivíduos dentro de um mesmo grupo ou família.

3.1 Modelo de fragilidade Multiplicativo

O termo fragilidade foi introduzido pela primeira vez na análise de sobrevivência por Vaupel *et al.* (1979), a fim de permitir a heterogeneidade não observada pela inclusão de um efeito aleatório, ou seja, de uma variável aleatória não observável, que representa as informações que não podem ou que não foram observadas. A fragilidade pode ser inserida no modelo de forma aditiva ou multiplicativa com o objetivo de avaliar a heterogeneidade entre as unidades na função de risco.

O modelo de fragilidade multiplicativo é uma extensão do modelo de Cox (1972), onde o risco individual depende de uma variável não observada V , a qual age multiplicativamente sobre a função de risco básico. Assim, o risco individual para o i-ésimo indivíduo no tempo t é dado por:

$$\lambda_M(t|v_i, x_i) = v_i\lambda_0(t)exp\left\{\mathbf{x_i'}\boldsymbol{\beta}\right\} \tag{3.1}$$

onde:

- v_i: representa a fragilidade do i-ésimo indivíduo;
- λ_0: representa a função de risco básico;
- $\boldsymbol{\beta}$: vetor de coeficientes a serem estimados;
- $\mathbf{x_i'}$: as covariáveis associadas ao i-ésimo indivíduo.

O modelo de fragilidade (3.1) assume estrutura de risco proporcional condicionado ao efeito aleatório. Como v_i representa um valor da variável aleatória não observável V, o risco individual cresce quando $v_i > 1$, decresce se $v_i < 1$ e para $v_i = 1$ o modelo de fragilidade (3.1) se reduz ao modelo de risco proporcional de Cox (Cox, 1972). O fato de a variável de fragilidade atuar de forma multiplicativa na função de risco implica de forma que quanto maior for o valor da variável de fragilidade, maior será a chance de ocorrer

a falha. Dessa forma, quanto maior for v_i, mais "frágeis" as observações pertencentes ao indivíduo i estão para falhar, daí o nome de fragilidade. Portanto, é esperado que o evento de interesse ocorra para os indivíduos mais "frágeis".

No contexto de modelo de riscos proporcionais, segundo Elbers e Ridder (1982) quando se trabalha com fragilidade é necessário que a distribuição do efeito aleatório tenha média finita para o modelo ser identificável.

A função de risco, sem covariáveis, para o i-ésimo indivíduo é dado por:

$$\lambda_M(t|v_i) = v_i\lambda_0(t), \tag{3.2}$$

com função de sobrevivência:

$$S_M(t|v_i) = [S_0(t)]^{v_i}, \tag{3.3}$$

em que $S_0(t)$ é a função de sobrevivência comum à população, representando a probabilidade do indivíduo estar vivo no tempo t, dado o efeito aleatório v.

Para obter a função de verossimilhança, é necessário encontrar a função de sobrevivência não condicional, assim é preciso integrar o termo de fragilidade, isto é:

$$S_M(t) = \int_0^\infty S_M(t|v)f_V(v)dv = \int_0^\infty [S_0t]^v f_V(v)dv = \int_0^\infty e^{-\Lambda_0(t)v}f_V(v)dv = L_V\left[\Lambda_0(t)\right] \tag{3.4}$$

em que $f_V(v)$ é a função densidade de probabilidade da variável de fragilidade e $L_V[\Lambda_0(t)]$ denota a transformada de Laplace aplicada na função de risco acumulada, $\Lambda_0(t)$ (WIENKE, 2010).

3.1.1 Distribuição de Fragilidade Gama

Considerando que a variável aleatória V, que representa a fragilidade, segue uma distribuição $V \sim Gama(\alpha, \alpha)$ dada em (2.19). A variância da variável de fragilidade, neste caso $Var(V) = 1/\alpha$ quantifica a heterogeneidade não observável entre os indivíduos. Quando α é suficientemente grande, há pouca variabilidade entre os mesmos, ou seja, uma população mais homogênea, assim os valores das variáveis de fragilidade serão aproximadamente iguais a 1, e consequentemente, a distribuição gama fica praticamente de-

generada no ponto 1 e, com isso, tem-se o modelo de riscos proporcionais de Cox para dados independentes. Por outro lado, um valor pequeno de α indica que há uma grande heterogeneidade não observável entre os indivíduos.

Assim, sem considerar covariáveis observadas, a função de sobrevivência e risco não condicional obtidas através da transformada de Laplace são dadas por:

$$S_M(t) = \left[1 + \frac{\Lambda_0(t)}{\alpha}\right]^{-\alpha} \text{ e } \lambda_M(t) = \frac{\alpha\lambda_0(t)}{\alpha + \Lambda_0(t)}. \tag{3.5}$$

E na presença de covariáveis:

$$S_M(t|\mathbf{x}) = \left[1 + \frac{\Lambda_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})}{\alpha}\right]^{-\alpha} \tag{3.6}$$

$$\lambda_M(t|\mathbf{x}) = \frac{\alpha\lambda_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})}{\alpha + \Lambda_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})}. \tag{3.7}$$

Para as funções de risco básico $\lambda_0(t)$ e risco base acumulado $\Lambda_0(t)$ podem ser atribuídas distribuições utilizadas para representar tempos de vida, como a Exponecial, Log-Normal, Gompertz e etc. (WIENKE, 2010).

3.1.2 Modelo de fragilidade multiplicativo exponencial

Para o caso em que a distribuição do tempo de sobrevivência é Exponencial(θ), cuja função densidade é dada em (2.12), a função de risco acumulada é dada por

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\, du = \theta t. \tag{3.8}$$

Então, utilizando (3.8), a função de risco não condicional e a função de sobrevivência não condicional são dadas, respectivamente, por

$$\lambda_M(t) = \frac{\theta}{1 + \frac{\theta t}{\alpha}} \tag{3.9}$$

e

$$S_M(t) = \left[1 + \frac{\theta t}{\alpha}\right]^{-\alpha}, \tag{3.10}$$

sendo θ a média do risco individual e $1/\alpha$ a variância da variável de fragilidade.

Estas funções correspondem às funções da distribuição de Pareto de segunda espécie ou distribuição Lomax (Lomax, 1954) com os parâmetros $(\delta, \frac{\nu}{\delta})$, cuja função densidade de probabilidade é dada por

$$f_L(t) = \frac{\nu}{\left(1 + \frac{\nu t}{\delta}\right)^{\delta+1}}, \tag{3.11}$$

em que $t > 0$, $\nu > 0$ e $\delta > 0$.

Seja $(t_1, t_2, \ldots, t_n)$ uma amostra aleatória de tempos de vida. Considerando (3.10) e (3.9) a função de verossimilhança é dada por

$$L(\nu, \delta) = \prod_{i=1}^{n} \lambda\left(t_i\right) S\left(t_i\right) = \prod_{i=1}^{n} \frac{\nu}{\left(1 + \frac{\nu t_i}{\delta}\right)^{\delta+1}}. \tag{3.12}$$

O logaritimo da função de verossimilhança é dado por

$$l(\nu, \delta) = n \log(\nu) - (\delta + 1) \sum_{i=1}^{n} \log\left(1 + \frac{\nu t_i}{\delta}\right). \tag{3.13}$$

As estimativas de máxima versossimlhança podem ser encontradas via maximização direta do logaritmo da função de verossimilhança ou via métodos numéricos como por exemplo método de Newton Raphson. Intervalos de confiança e testes de hipóteses para os parâmetros podem ser realizados, considerando-se a distribuição normal assintótica para os estimadores de máxima verossimilhança (Lawless, 2011).

3.2 Modelo de fragilidade aditivo

Rocha (1995) propôs um modelo de fragilidade linear, assumindo, para a função de

risco no instante t de uma unidade com fragilidade v, a função de risco com a estrutura aditiva, esta que é dada por:

$$\lambda_A(t|v) = \lambda_0(t) + v, \;\; t > 0, v > 0, \tag{3.14}$$

em que v é uma variável de fragilidade não negativa e $\lambda_0(t)$ uma função do tempo comum a todos os indivíduos, que é chamada função de risco básico para uma unidade padrão com fragilidade nula ($v = 0$).

A função de sobrevivência condicionada à variável de fragilidade v pode ser obtida pela relação com a função de risco acumulada $\Lambda(t|v)$, que por sua vez pode ser obtida por $\Lambda(t|v) = \int_0^t \lambda(u|v)du$.

Assim, do modelo de risco com fragilidade aditiva dado em (3.14) tem-se que a função de risco acumulada é dada por:

$$\Lambda_A(t|v) = \int_0^t \lambda_A(u|v)du = \int_0^t (\lambda_0(u) + v)du = vt + \int_0^t \lambda_0(u)du = vt + \Lambda_0(t).$$

Dessa forma, a função de sobrevivência condicionada é dada por:

$$S_A(t|v) = \exp\{-\Lambda(t|v)\} = \exp\{-vt - \Lambda_0(t)\}, \tag{3.15}$$

em que $\Lambda_0(t)$ é a função de risco básico acumulada.

Para obter a função de sobrevivência não condicional $S_A(t)$ será utilizada a transformada de Laplace da seguinte forma:

$$S_A(t) = \int_0^\infty S_A(t|v) f_V(v)\mathrm{d}v, \tag{3.16}$$

em que $f_V(v)$ é a função densidade de probabilidade da variável de fragilidade v e $S_A(t|v)$ é a função de sobrevivência condicional (3.15).

$$\begin{aligned} S_A(t) &= \int_0^\infty S_A(t|v) f_V(v)\mathrm{d}v = \int_0^\infty \exp\{-vt - \Lambda_o(t)\} f_V(v)\mathrm{d}v \\ &= \exp\{-\Lambda_0(t)\} \int_0^\infty \exp\{-vt\} f_V(v)dv \\ &= \exp\{-\Lambda_0(t)\} L_V(t), \end{aligned} \tag{3.17}$$

em que $L_V(t)$ é a de transformada da Laplace aplicada no tempo.

A função de risco marginal $\lambda(t)$ é dada pela seguinte relação com $S_A(t)$ dada em (3.17):

$$\begin{aligned}\lambda_A(t) &= -\frac{d}{dt}[log(S_A(t)] = -\frac{d}{dt}[log(\exp\{-\Lambda_0(t)\}L_V(t))] \\ &= -\frac{d}{dt}[-\Lambda_0(t) + log(L_V(t))] = \frac{d}{dt}[\Lambda_0(t)] - \frac{d}{dt}[log(L_V(t))] \\ &= \lambda_0(t) - \frac{L_V^{'}(t)}{L_V(t)}, \end{aligned} \tag{3.18}$$

em que $L_V^{'}(t)$ é a derivada da transforma da Laplace em relação a t.

Ainda pode-se encontrar a distribuição condicional da variável de fragilidade V dado $T > t$, ou seja, a distribuição da fragilidade dos indivíduos que sobreviveram até o tempo t. Note que é possível escrever a função de densidade de probabilidade $f_V(v|T > t)$ em termos da função de sobrevivência condicional dada em (3.15), da função de sobrevivência maginal dada em (3.17) e da densidade da variável de fragilidade $f_V(v)$, da seguinte forma

$$f_V(v|T > t) = \frac{S_V(t|v)f_V(v)}{S_V(t)} = \frac{exp\{-vt - \Lambda_0(t)\}f_V(v)}{exp\{-\Lambda_0(t)\}L_V(t)} = \frac{exp\{-vt\}f_V(v)}{L_V(t)}. \tag{3.19}$$

Podemos obter ainda a distribuição da fragilidade V dado $T = t$ a partir da seguinte expressão

$$\begin{aligned} f(v|T=t) &= \frac{f(t|v)f(v)}{f(t)} = \frac{S(t|v)\lambda(t|v)f(v)}{S(t)\lambda(t)} \\ &= \frac{(\lambda_0(t) + \mathbf{x}_i'\beta)}{\left((1+\alpha t)^{-1} + \lambda_0(t) + \mathbf{x}_i'\beta\right)} \frac{\left(\frac{1}{\alpha}+t\right)^{\frac{1}{\alpha}}}{\Gamma(1/\alpha)} \exp\left\{-v\left(\frac{1}{\alpha}+t\right)\right\} v^{\frac{1}{\alpha}-1} \\ &\quad + \frac{(1+\alpha t)^{-1}}{\left((1+\alpha t)^{-1} + \lambda_0(t) + \mathbf{x}_i'\beta\right)} \frac{\left(\frac{1}{\alpha}+t\right)^{\frac{1}{\alpha}+1}}{\Gamma(1/\alpha+1)} \exp\left\{-v\left(\frac{1}{\alpha}+t\right)\right\} v^{\frac{1}{\alpha}+1-1} \\ &= w_1 \times G_1 + w_2 \times G_2. \end{aligned} \tag{3.20}$$

Em que G_1 é a função de densidade de uma distribuição Gama$(1/\alpha, 1/\alpha + t)$ e G_2 é a função de densidade de uma distribuição Gama$(1/\alpha + 1, 1/\alpha + t)$, com $w_1 + w_2 = 1$. O

valor esperado de $V|T=t$ pode ser calculado a partir de (3.20) da seguinte maneira

$$\begin{aligned} E[V|T=t] &= \int_0^\infty v f(v|T=t)\mathrm{d}v = \int_0^\infty v(w_1 \times G_1 + w_2 \times G_2)\mathrm{d}v \\ &= \int_0^\infty v \times w_1 \times G_1 \mathrm{d}v + \int_0^\infty v \times w_2 \times G_2 \mathrm{d}v \\ &= w_1 \int_0^\infty v \times G_1 \mathrm{d}v + w_2 \int_0^\infty v \times G_2 \mathrm{d}v \\ &= w_1 \left(\frac{1}{1+\alpha t}\right) + w_2 \left(\frac{\alpha+1}{1+\alpha t}\right). \end{aligned} \tag{3.21}$$

3.2.1 Modelo aditivo com fragilidade Gama

A variável de aleatória V que representa a fragilidade segue uma distribuição $Gama(\alpha, \alpha)$ dada em (2.19). A variância da variável de fragilidade, neste caso $Var(V) = 1/\alpha$ quantifica a heterogeneidade não observável entre os indivíduos.

Considerando a distribuição Gama como núcleo de uma distribuição sua transformada de Laplace é dada por:

$$L_V(t) = \left(\frac{\alpha}{t+\alpha}\right)^\alpha.$$

Dessa forma substituindo $L_V(t)$ em 3.17 a função de sobrevivência não condicionada é dada por:

$$S_A(t) = \left(\frac{\alpha}{t+\alpha}\right)^\alpha \exp\{-\Lambda_0(t)\}. \tag{3.22}$$

Logo,

$$\lambda_A(t) = \frac{\alpha}{t+\alpha} + \lambda_0(t). \tag{3.23}$$

Pela Equação (3.19) tem-se que para $V \sim Gama(\alpha, \alpha)$,

$$\begin{aligned} f_V(v|T>t) &= exp\{-vt\}\frac{\alpha^\alpha}{\Gamma(\alpha)} exp\left\{-v\alpha\right\} v^{\alpha-1} \left(\frac{t+\alpha}{\alpha}\right)^\alpha \\ &= \frac{(t+\alpha)^\alpha}{\Gamma(\alpha)} exp\left\{-v(t+\alpha)\right\} v^{\alpha-1}, \end{aligned} \tag{3.24}$$

ou seja, $V|T>t \sim Gama\,(\alpha, t+\alpha)$ e portanto $E[V|T>t] = \alpha/(t+\alpha)$ e $Var[V|T>t] =$

$\alpha/(t+\alpha)^2$.

3.2.2 Modelo de fragilidade aditivo Gama Exponencial e Weibull

É possível escolher diversas distribuições para o risco de base, dentre elas a distribuição Weibull(θ, γ) com função de densidade definida em (2.15), função de risco definida em (2.18) e função de risco acumulado $\Lambda_W(t) = \theta t^\gamma$. Quando $\gamma = 1$ tem-se a distribuição Exponencial(θ) com $\lambda_E(t) = \theta$ e $\Lambda_E(t) = \theta t$. Nesse caso, considerando que a fragilidade segue uma $Gama(\alpha, \alpha)$, a função de sobrevivência não condicional dada em (3.22) fica da forma:

$$S_A(t) = \left(\frac{\alpha}{t+\alpha}\right)^\alpha exp[-\theta t],$$

e a função de risco marginal é expressa como:

$$\lambda_A(t) = \frac{\alpha}{t+\alpha} + \theta.$$

A partir das funções $S_A(t)$ e $\lambda_A(t)$ é possível obter a função de verossimilhança e assim estimar os parâmetros via método de MV.

Considerando agora que os tempos de vida dos indivíduos em risco seguem distribuição $Weibull(\theta, \gamma)$ e que a fragilidade segue uma $Gama(\alpha, \alpha)$, dada em (2.19), tem-se que a função de risco (3.23) é definida por:

$$\lambda_A(t) = \frac{\alpha}{t+\alpha} + \gamma\theta t^{\gamma-1} \tag{3.25}$$

e a função de sobrevivência (3.22) por:

$$S_A(t) = \left(\frac{\alpha}{t+\alpha}\right)^\alpha \exp[-\theta t^\gamma]. \tag{3.26}$$

A função de verossimilhança para fragilidade $Gama(\alpha, \alpha)$ e risco de base $\lambda_0(t)$ é dada por:

$$L(\pi,\alpha,\beta|\mathcal{D}) = \prod_{i=1}^{n}\left\{\left[\frac{\alpha}{t_i+\alpha}+\lambda_0(t_i)\right]^{\delta_i}\left(\frac{\alpha}{t_i+\alpha}\right)^{\alpha} exp[-\Lambda_0(t_i)]\right\}, \tag{3.27}$$

em que $\delta_i = 1$ se t_i for um tempo de falha e é zero caso contrário, $\mathcal{D} = \{(t_i,\delta_i)\}_{i=1}^{n}$ e π é o vetor de parâmetros da distribuição de base. O logaritmo da função de verossimilhança em (3.27) denotado por $l(\pi,\alpha,\beta|\mathcal{D}) = log[L(\pi,\alpha,\beta|\mathcal{D})]$ é dado por:

$$l(\pi,\alpha,\beta|\mathcal{D}) = \sum_{i=1}^{n}\delta_i log\left[\frac{\alpha}{t_i+\alpha}+\lambda_0(t_i)+\right] + \alpha\sum_{i=1}^{n} log\left[\frac{\alpha}{t_i+\alpha}\right] - \sum_{i=1}^{n}\Lambda_0(t_i)$$

A função de risco de base $\lambda_0(.)$ e risco acumulada $\Lambda_0(.)$ podem assumir diferentes distribuiçãoes tais como: exponencial, Weibull, gama entre outras. As estimativas dos parâmetros são obtidas numericamente.

3.3 Aplicação para Dados de Leucemia

Os dados dessa aplicação foram estudados inicialmente por Feigl e Zelen (1965) e podem ser visualizados em Colosimo e Giolo (2006). Esses dados são de um estudo realizado com pacientes que tinham leucemia aguda, e foram observados semanalmente. Dentre as covariáveis observadas está a variável indicadora da ausência do antígeno Calla na superfície dos blatos, que indica a presença de linfomas e carcinomas, denotada por (Ag-), ou seja, $x = 1$ para o grupo (Ag-) e $x = 0$ para o grupo (Ag+) que apresentaram o antígeno Calla na superfície dos blatos. O grupo Ag+ tinha 17 pacientes com tempo mediano de sobrevivência de 56 semenas com $IC(95\%) = (22; 121)$ e o grupo Ag- contava com 16 pacientes com apenas 7,5 semanas de tempo mediano de sobrevivência.

Tabela 3.1: Estatísticas descritivas.

Grupos	n	Tempo mediano	LI (95%)	LS (95%)
Ag+	17	56	22	121
Ag-	16	7,5	4	43

De acordo com Tabela 3.1 o grupo Ag+ apresentou um tempo mediano de sobrevivência de 56 semanas, enquanto o grupo Ag- apresentou apenas 7,5 semanas de tempo

mediano de sobrevivência. Vale ressaltar que LI é o limite inferior do intervalo de confiança e que LS é o limite superior desse.

Na Figura 3.1 percebe-se que as curvas de sobrevivência estimadas pelo estimador de Kaplan e Meier (1958) (K-M) parecem indicar que os pacientes do grupo Ag+ tem maior sobrevida que os pacientes do grupo Ag-. Para verificar se existe uma diferença entre os tempos de sobrevida, o teste não paramétrico de Log-Rank (Mantel e Haenszel, 1959) foi aplicado e obteve-se um p-valor$=0.004$, de onde conclui-se que, ao nível de significância de 5%, que há evidências estatísticas de diferença entre os tempos de sobrevivências dos pacientes pertencentes aos Ag- e Ag+.

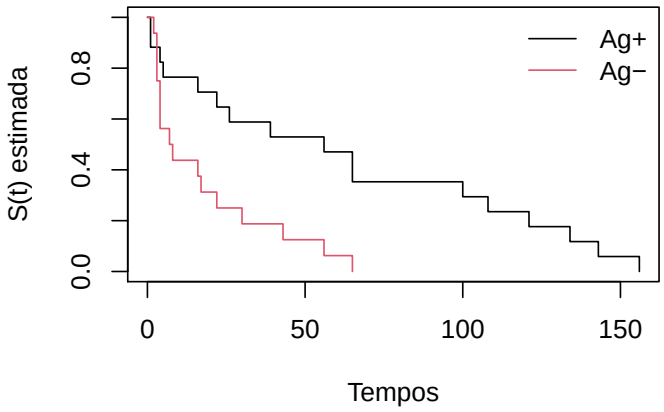

Figura 3.1: Curvas de sobrevivência K-M estratificada por grupo.

Ajuste do Modelo de fragilidade aditivo Gama-Exponencial

Considerando a distribuição de base expenencial tem-se que na Tabela 3.2 os resultados das estimativas de parâmetros, erro padrão e seus intervalos de confiança de 95% para o ajuste do modelo de fragilidade aditiva gama-Exponencial a partir da função de verossimilhança dada na Equação 3.27.

Tabela 3.2: Estimativas de máxima verossimilhança (EMV), erro padrão (EP), intervalo de confiança (IC(95%)) para os parâmetros do modelo (3.25).

Parâmetros	EMV	EP	LI (95%)	LS (95%)
β	0,0464	0,0166	0,0190	0,0737
α	0,0124	0,0512	0,0000	0,0969
μ	0,0138	0,0053	0,0050	0,0226

Pelos resultados apresentadas na Tabela 3.2 observa-se que $\hat{\beta} = 0,0464$ e o respectivo intervalo de confiança para o parâmetro β não inclui o valor zero, então a covariável "grupo"é signifivativa, ou seja, ser do grupo Ag- aumenta o risco de morte em pacientes com leucemia aguda. Além disso, o valor estimado da variância que representa a hetorogeneidade não observada, $\hat{Var}(V) = 1/\hat{\alpha} = 1/0,0124 = 80,65$, ou seja, há uma alta heterogeneidade entre os indivíduos e há evidências de que existem outros fatores não observados que podem ter efeito significativo no tempo de vida dos pacientes com leucemia aguda grave.

Na Figura 3.2 observa-se que o modelo aditivo com fragilidade $Gama(\alpha, \alpha)$ e função de risco de base $Exponencial(\theta)$ parece se ajustar bem aos dados, pois se aproxima da curva estimada não parametricamente pelo estimador K-M.

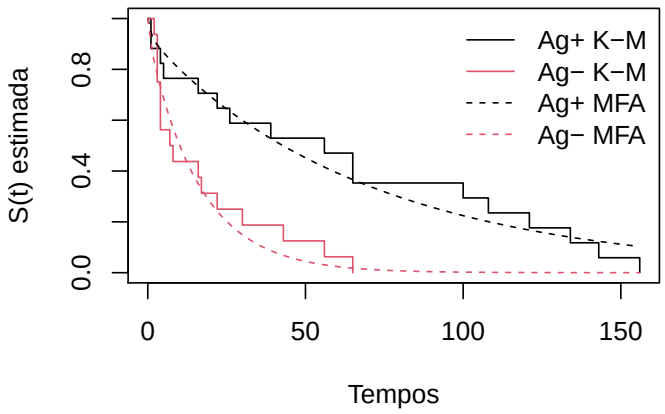

Figura 3.2: Comparação das curvas de sobrevivências estimadas pelo modelo de fragilidade aditiva (MFA) com as curvas estimadas pelo estimador K-M.

Pela Equação 3.24 pode-se obter o valor médio da fragilidade para os indivíduos que sobreviveram até o tempo t. Por exemplo, para os sobreviventes na semana 50 a fragilidade média foi de 0,0003, enquanto na semana 10 a fragilidade média dos sobreviventes foi de 0,0054, ou seja, à medida que o tempo passa os indivíduos que permanecem no estudo são em média menos frágeis.

Tabela 3.3: Fragilidade individual estimada dos 33 pacientes com leucemia aguda divididos por grupo.

i	Tempos (Grupo Ag+)	v_i	i	Tempos (Grupo Ag-)	v_i
1	65	0,0004	18	56	0,0005
2	156	0,0001	19	65	0,0004
3	100	0,0002	20	17	0,0037
4	134	0,0001	21	7	0,0180
5	16	0,0041	22	16	0,0041
6	108	0,0002	23	22	0,0023
7	121	0,0002	24	3	0,0804
8	4	0,0487	25	4	0,0487
9	39	0,0009	26	2	0,1596
10	143	0,0001	27	3	0,0804
11	56	0,0005	28	8	0,0141
12	26	0,0018	29	4	0,0487
13	22	0,0023	30	3	0,0804
14	1	0,4767	31	30	0,0014
15	1	0,4767	32	4	0,0487
16	5	0,0328	33	43	0,0008
17	65	0,0004	-	-	-

Para calcular as fragilidades individuais apresentadas na Tabela 3.3 foram utilizadas as estimativas da Tabela 3.2, ou seja, aplicou-se o valor estimado de α na Equação 3.24, processo também chamado de *plug-in*. Observando as fragilidades individuais estimadas apresentadas na Tabela 3.3, percebe-se que em geral os inidíduos mais frágeis estão no Grupo Ag-, isto é, os indivíduos que não apresentaram o antígeno Calla na superfície dos blastos têm maior risco de morte do que os indivíduos do Grupo Ag+.

3.4 Conclusão

Este capítulo teve o intuito de apresentar os conceitos do modelo de fragilidade, caracterizado por utilizar uma variável aleatória não observável como hipótese principal. Essa fragilidade pode ser introduzida de forma multiplicativa ou aditiva na função de risco, e portanto, foram abordados nesse capítulo essas duas formas. Foi também realizada uma breve aplicação para o modelo de risco aditivo, utilizando dados de pacientes que tinham leucemia aguda. Para dar continuidade a este trabalho, no próximo capítulo será abordado o modelo de mistura padrão com fragilidade aditiva.

Capítulo 4

Modelo de mistura padrão com fragilidade aditiva

Os modelos de cura assumem implicitamente que os indivíduos que sofreram o evento de interesse possuem riscos homogêneos. No entanto, uma forma de medir a heterogeneidade observada é adicionando covariáveis ao modelo. Assim, uma parcela da heterogeneidade pode ser explicada por covariáveis, embora, exista um grau de heterogeneidade induzida por fatores de riscos não observáveis. Os modelos que incorporam a heterogeneidade não observável entre os indivíduos são conhecidos como modelos de fragilidade (Vaupel *et al.*, 1979).

Gonzales *et al.* (2013) propuseram o modelo de mistura padrão com fragilidade multiplicativa. Neste trabalho será proposto o modelo de Mistura padrão com fragilidade Aditiva (ver seção 2.4).

4.1 Modelos de Longa duração

Em modelos padrões de sobrevivência, supõem-se de que todas as unidades envolvidas no experimento atingirão o evento de interesse. No entanto, considerar tal modelo para determinados conjuntos de dados pode não ser adequado. Existem dados de sobrevivência em que uma parcela das unidades em estudo nunca apresentará o evento de interesse, mesmo se acompanhados por um tempo suficientemente grande. Por exemplo, uma lâmpada cedo ou tarde falhará, porém, um paciente "curado" de câncer pode nunca vir apresentar a recorrência do tumor. Diz-se então, que esses indivíduos ou unidades são "imunes" ao evento de interesse e a população aos quais eles pertencem possui uma

fração de cura. Neste contexto, o modelo de mistura padrão proposto por Berkson e Gage (1952) é o mais conhecido para estimar a fração de curados.

Os modelos de longa duração possuem uma vantagem em relação aos modelos padrões de sobrevivência, no sentido de incorporarem a heterogeneidade de duas subpopulações. Dentre esses modelos, o tipo mais comum é o modelo de mistura, no qual se considera que a população é dividida em duas subpopulações (imunes e suscetíveis) ao evento de interesse.

4.1.1 Modelo de mistura padrão

O modelo de mistura padrão proposto por Berkson e Gage (1952) é um dos tipos mais comuns na análise de sobrevivência para ajustar dados de longa duração. Este consiste em uma mistura de distribuições paramétricas, sendo uma função de sobrevivência imprópria considerada para a população total (curados e não curados), chamada de sobrevivência populacional ($S_{pop}(t)$) e uma função de sobrevivência própria para a parte da população formada pelos não curados.

Desta forma, considerando T uma variável aleatória não negativa e contínua, representando o tempo de vida, sabe-se que

$$P(T > t|M_i = 1) = S(t) \; e \; P(T > t|M_i = 0) = 1, \tag{4.1}$$

em que, $M_i = 0$ se o indivíduo i não está em risco e $M_i = 1$ se o indivíduo i está em risco, com $i = 1, \ldots, n$. Assumindo que $P(M_i = 0) = p_0$ e $P(M_i = 1) = 1 - p_0$, sendo $p_0 \in (0, 1)$ a proporção de curados ou imunes, a probabilidade de o tempo de vida ser maior que um determinado tempo t, independente do grupo a que ele pertença é dada por

$$\begin{aligned} S_{pop}(t) &= P(T > t) = P(T > t|M_i = 0)P(M_i = 0) + P(T > t|M_i = 1)P(M_i = 1) \\ &= p_0 + (1 - p_0)S(t), \quad t \geq 0. \end{aligned} \tag{4.2}$$

Assim, a função de sobrevivência populacional é dada por

$$S_{pop}(t) = p_0 + (1 - p_0)S(t), \quad t \geq 0, \tag{4.3}$$

em que $S(.)$ representa a função de sobrevivência própria associada aos indivíduos em

risco e p_0 a proporção de curados.

A função de sobrevivência $S_{pop}(t)$ possui as seguintes propriedades:

- Se $p_0 = 0$, então $S_{pop}(t) = S(t)$;
- $S_{pop}(0) = 1$;
- $S_{pop}(t)$ é decrescente;
- $\lim_{t\to\infty} S_{pop}(t) = p_0$.

A última propriedade retrata o fato da função de sobrevivência populacional ser imprópria, pois a curva de sobrevivência estabiliza em p_0 (proporção dos indivíduos não suscetíveis ao evento de interesse), justamente a probabilidade de cura da população.

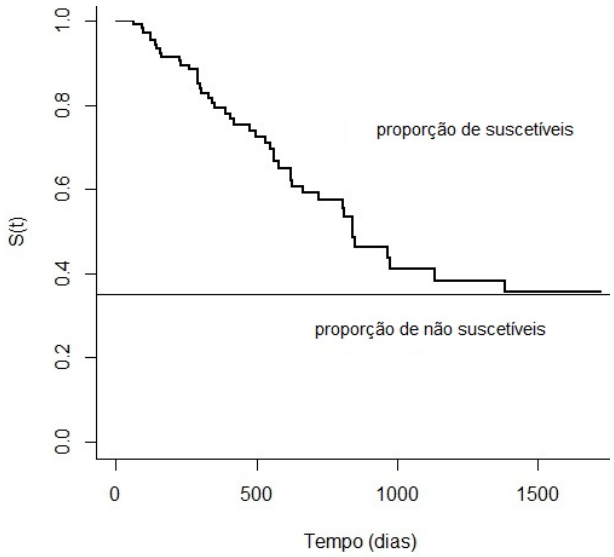

Figura 4.1: Curva de sobrevivência do modelo de fração de cura.

A função de densidade imprópria é

$$f_{pop}(t) = -\frac{dS_{pop}(t)}{dt} = (1 - p_0)f(t), \tag{4.4}$$

em que $f(.)$ representa a função de densidade própria relativa ao grupo dos indivíduos em risco. Com função de risco populacional dada por

$$\lambda_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} = \frac{(1 - p_0)f(t)}{p_0 + (1 - p_0)S(t)}. \tag{4.5}$$

Da equação anterior, a função de risco própria é dada por

$$\lambda(t) = \frac{S_{pop}(t)\lambda_{pop}(t)}{(1-p_0)S(t)} = \left[\frac{S_{pop}(t)}{S_{pop}(t)-(1-p_0)}\right]\lambda_{pop}(t). \tag{4.6}$$

Como $\{S_{pop}(t)/\left[S_{pop}(t)-(1-p_0)\right]\} > 1$, tem-se que $\lambda_{pop}(t) < \lambda(t)$, ou seja, a função de risco da população é limitada pela função de risco de base. De (4.6) decorre que $\lambda(t)$ não possui a propriedade de riscos proporcionais, uma vez que $\{S_{pop}(t)/\left[S_{pop}(t)-(1-p_0)\right]\}$ sempre dependerá de t. Observe ainda que

$$\lim_{t\to\infty}\lambda_{pop}(t) = \lim_{t\to\infty}\frac{(1-p_0)f(t)}{S_{pop}(t)} = \left(\frac{1-p_0}{p_0}\right)\lim_{t\to\infty}f(t) = 0. \tag{4.7}$$

O resultado em (4.7) revela que conforme o tempo aumenta, o risco da população converge para o valor zero, indicando que a curva de sobrevivência populacional estabilizar em um determinado valor (fração de cura), e que uma parcela dos indivíduos em estudo não falhou e possivelmente eles foram curados durante o experimento.

Considerando que $p_0 \in (0,1)$ pode ser explicada pelas covariáveis observadas, a função de ligação logit será considerada neste trabalho, mas outras funções de ligação como probit e complementar log-log podem também ser consideradas. A fração de cura na presença de covariáveis considerando a função de ligação logit é expressa por

$$p_0(\mathbf{x}) = \frac{\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)}, \tag{4.8}$$

em que $\mathbf{x}' = (1, x_1, \ldots, x_q)$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)^\top$ são o conjunto de covariáveis e seus respectivos coeficientes da regressão.

A função de sobrevivência populacional na presença de covariáveis pode ser reescrita por

$$S_{pop}(t \mid \mathbf{x}) = p_0(\mathbf{x}) + \left[1-p_0(\mathbf{x})\right]S(t|\mathbf{x}),$$

em que $S(t|\mathbf{x})$ é a função de sobrevivência dos pacientes não imunes.

A fim de estimar os parâmetros do modelo, utilizou-se o método de máxima verossimilhança. Devido à complexidade da função de verossimilhança, os estimadores de máxima verossimilhança não possuem uma expressão fechada, e por isso a estimação dos parâmetros deve ser obtida através de algum procedimento numérico, como os implementados no software R .

4.1.2 Modelo de mistura padrão com fragilidade aditiva

Para o modelo de fragilidade aditivo geral expresso em (3.17) verifica-se que quando o tempo é muito grande a função de sobrevivência não condicional tende a zero, não captando a fração de cura. Sendo assim, torna-se interessante estudar um modelo aditivo que consiga estimar a proporção de curados, quando esta existir. Este é o modelo de mistura padrão com fragilidade aditiva que pode ser escrito a partir da Equação (4.3) substituindo $S(t)$ pela Equação (3.17), ou seja, esse modelo segue a seguinte equação:

$$S_{pop}(t) = p_0 + (1 - p_0)\exp\{-\Lambda_0(t)\}L_V(t) \tag{4.9}$$

em que $\Lambda_0(t)$ é a função de risco base acumulada, que depende da escolha de distribuições paramétricas tais como: Exponencial, Weibull, Gompertz entre outras, e $L_V(t)$ é a transformada de Laplace do tempo que depende da escolha para a distribuição da variável de fragilidade.

Pela relação da função de densidade de probabilidade com a função de sobrevivência, $f(t) = -\frac{d}{dt}S(t)$, utilizando a Equação 4.3 obtem-se que

$$f_{pop}(t) = -\frac{d}{dt}S_{pop}(t) = (1 - p_0)f(t),$$

em que $f(t)$ é a função densidade de probabilidade dos indivíduos que estão em risco. Com esse resultado pode-se reescrever $f_{pop}(t)$ da seguinte forma:

$$f_{pop}(t) = (1 - p_0)f(t) = (1 - p_0)\lambda(t)S(t) \quad t \geq 0. \tag{4.10}$$

Para o modelo de fragilidade aditivo dado em (3.17) tem-se que (4.10) pode ser reecrita da seguinte forma

$$f_{pop}(t) = (1 - p_0)f(t) = (1 - p_0)\left(\lambda_0(t) - \frac{L'_V(t)}{L_V(t)}\right)\exp\{-\Lambda_0(t)\}L_V(t) \quad t \geq 0. \tag{4.11}$$

4.1.3 Inferência

Para determinar a função de verossimilhança, considera-se para a i-ésima observação os dados observados $\mathcal{D} = \{(t_i, \delta_i)\}_{i=1}^n$, em que t_i denota o tempo observado, δ_i é a variável indicadora de censura, com $\delta_i = 1$ se t_i é não censurado e $\delta_i = 0$ caso contrário. Assim, a

função de verossimilhança baseada nos dados observados e utilizando as equações (4.11) e (4.9) é dada por

$$\begin{aligned} L(\pi|\mathcal{D}) &= \prod_{i=1}^{n}[f_{pop}(t_i)]^{\delta_i}[S_{pop}(t_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n}[(1-p_0)\lambda(t_i)S(t_i)]^{\delta_i} \times [p_0+(1-p_0)S(t_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n}\left[(1-p_0)\left(\lambda_0(t_i)-\frac{L_V'(t_i)}{L_V(t_i)}\right)\exp\{-\Lambda_0(t_i)\}L_V(t)\right]^{\delta_i} \\ &\quad \times[p_0+(1-p_0)\exp\{-\Lambda_0(t_i)\}L_V(t_i)]^{1-\delta_i} \end{aligned} \tag{4.12}$$

A estimação do vetor de parâmetros do modelo, π, será feita através da maximização do logaritmo da função de verossimilhança, $l(\pi|\mathcal{D}) = log(L(\pi|\mathcal{D}))$. Para estimar o parâmetro da proporção de cura (p_0) utilizou-se a função de ligação logito e para estimá-lo é necessário substituir os parâmetros da regressão por suas estimativas obtidas na função. Foram considerados desvios padrão assintóticos das estimativas dos parâmetros, obtidos através da inversão da matriz de informação observada.

Como V é uma variável aleatória, é possível considerar diferentes distribuições de probabilidade, por exemplo a distribuição Gama, Log-Normal, Gaussiana Inversa, entre outras. Características gerais das distribuições para o termo da fragilidade foram estudadas por Hougaard (1995). Neste trabalho será considerada a distribuição gama para o termo da fragilidade. A vantagem de considerar a distribuição gama se deve à forma fechada da transformada de Laplace resultando em uma expressão analítica tratável para a função de sobrevivência não condicional.

4.2 Modelo Aditivo de fragilidade Gama

Assumindo que a variável de fragilidade segue uma distribuição Gama e substituindo a função de sobrevivência não condicional (3.22) no modelo de mistura padrão (4.3), tem-se que o modelo de mistura padrão com fragilidade aditivo é dado por:

$$S_{pop}(t) = p_0 + (1-p_0)\left(\frac{\alpha}{t+\alpha}\right)^{\alpha}\exp\{-\Lambda_0(t)\} \quad t \geq 0, \tag{4.13}$$

em que p_0 é a proporção de curados na população e $\Lambda_0(t)$ é a função de risco básico acumulada, que pode assumir diferentes distribuições tais como: Exponencial, Weibull, Gompertz, entre outras.

Assumindo que a função de risco de base vem de uma distriubuição Weibull tem-se:

$$S_{pop}(t) = p_0 + (1 - p_0)\left(\frac{\alpha}{t+\alpha}\right)^{\alpha} exp\{-\theta t^{\gamma}\} \quad t \geq 0. \tag{4.14}$$

Na Figura 4.2 observa-se o comportamento da função de sobrevivência dada em Equação (4.14) para diferentes valores dos parâmetros da distribuição de base e para diferentes frações de cura com o parâmetro da distribuição de fragilidade Gama fixado em $\alpha = 1$.

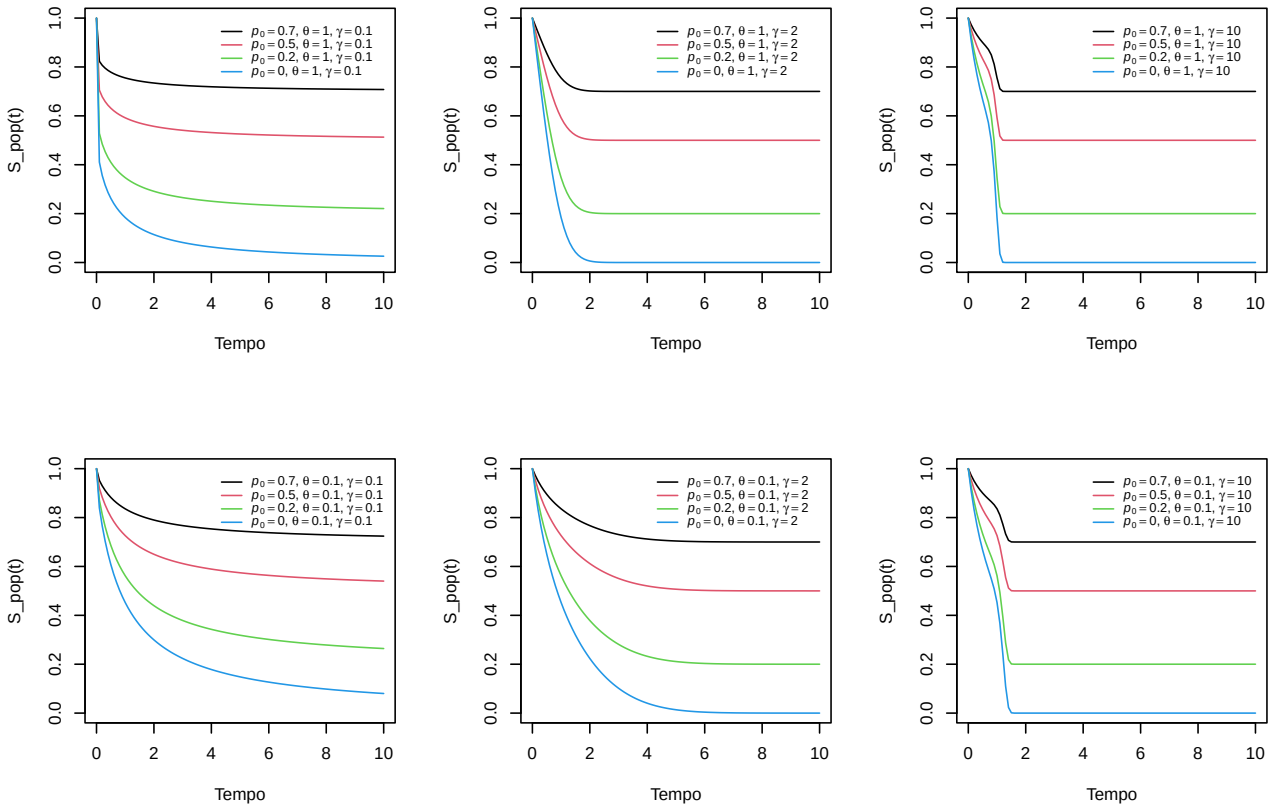

Figura 4.2: Curvas de sobrevivência do modelo 4.14 para o risco de base de uma distribuição Weibull e com o parâmetro da distribuição de fragilidade $\alpha = 1$.

Percebe-se que para valores de θ pequenos as curvas de sobrevivência caem mais

rapidamente para a proporção de curados. Variando o valor do parâmetro γ observa-se a flexibilidade desse modelo com risco de base de uma distribuição Weibull.

A função de verossimilhança do modelo de mistura padrão com fragilidade aditivo $Gama(\alpha, \alpha)$ (2.19) é expressa pela seguinte equação

$$\begin{aligned} L(\pi|\mathcal{D}) &= \prod_{i=1}^{n} \left[(1-p_0) \left(\lambda_0(t_i) + \frac{\alpha}{t_i+\alpha} \right) \exp\{-\Lambda_0(t_i)\} \left(\frac{\alpha}{t_i+\alpha} \right)^{\alpha} \right]^{\delta_i} \\ &\quad \times \left[p_0 + (1-p_0) \exp\{-\Lambda_0(t_i)\} \left(\frac{\alpha}{t_i+\alpha} \right)^{\alpha} \right]^{1-\delta_i} \end{aligned} \tag{4.15}$$

Para completar a modelagem paramétrica é nessário assumir uma distribuição para o tempo de sobrevivência dos indivíduos em risco e a partir dessa suposição obter uma forma paramétrica para a função de risco de base ($\lambda_0(t)$). Assumindo que essa segue a distribuição Weibull, com função de densidade definida em (2.15), função de risco definida em (2.18) e função de risco acumulado $\Lambda_W(t) = \theta t^{\gamma}$, tem-se:

$$\begin{aligned} L(\pi|\mathcal{D}) &= \prod_{i=1}^{n} \left[(1-p_0) \left(\gamma\theta t^{\gamma-1} + \frac{\alpha}{t_i+\alpha} \right) \exp\{-\theta t^{\gamma}\} \left(\frac{\alpha}{t_i+\alpha} \right)^{\alpha} \right]^{\delta_i} \\ &\quad \times \left[p_0 + (1-p_0) \exp\{-\theta t^{\gamma}\} \left(\frac{\alpha}{t_i+\alpha} \right)^{\alpha} \right]^{1-\delta_i} \end{aligned} \tag{4.16}$$

4.2.1 Modelo de mistura padrão com fragilidade aditiva na presença de covariáveis

Nessa modelagem as covariáveis podem ser introduzidas no parâmetro da fração de cura, ou na função de sobrevivência, ou ainda nos dois termos. As covariáveis que podem ter efeito na fração de cura e na função de sobrevivência podem ser as mesmas. Dessa forma, o modelo de mistura padrão com covariáveis nos dois termos pode ser escrito da seguinte forma

$$S_{pop}(t|\mathbf{x}) = p_0(\mathbf{x}) + (1 - p_0(\mathbf{x}))S(t|\mathbf{x}) \tag{4.17}$$

e o modelo de mistura padrão com fragilidade aditiva na presença de covariáveis pode ser expresso como

$$S_{pop}(t|\mathbf{x}) = p_0(\mathbf{x}) + (1 - p_0(\mathbf{x}))\exp\{-\Lambda_0(t) - \mathbf{x}'\boldsymbol{\beta}t\}L_V(t) \tag{4.18}$$

Para relacionar a fração de cura às covariáveis pode-se utilizar as funções de ligação logito, probito ou complementar log-log, visto que essas funções assumem valores no intervalo (0,1) e são usadas para modelar probabilidades.

Considerou-se para a i-ésima observação os dados observados $\mathcal{D} = (t_i, \delta_i, x_i)_{i=1}^n$, em que t_i denota o tempo observado, δ_i é a variável indicadora de censura, com $\delta_i = 1$ se t_i é não censurado e $\delta_i = 0$ caso contrário e $\mathbf{x}_i$ representa o vetor de covariáveis que tem efeito na distribuição latente e na taxa de cura. De maneira similar a Equação (4.12) obteve-se a seguinte função de verossimilhança

$$\begin{aligned}
L(\theta|\mathcal{D}) &= \prod_{i=1}^{n}[f_{pop}(t_i|\mathbf{x}_i)]^{\delta_i}[S_{pop}(t_i|\mathbf{x}_i)]^{1-\delta_i} \\
&= \prod_{i=1}^{n}[(1 - p_0(\mathbf{x}_i))\lambda(t_i|\mathbf{x}_i)S(t_i|\mathbf{x}_i)]^{\delta_i} \times [p_0(\mathbf{x}_i) + (1 - p_0(\mathbf{x}_i))S(t_i|\mathbf{x}_i)]^{1-\delta_i} \\
&= \prod_{i=1}^{n}\left[(1 - p_0(\mathbf{x}_i))\left(\lambda_0(t_i) + \mathbf{x}_i'\beta - \frac{L_V'(t_i)}{L_V(t_i)}\right)\exp\{-\Lambda_0(t_i) - \mathbf{x}_i'\beta t_i\}L_V(t_i)\right]^{\delta_i} \\
&\quad \times [p_0(\mathbf{x}_i) + (1 - p_0(\mathbf{x}_i))\exp\{-\Lambda_0(t_i) - \mathbf{x}_i'\beta t_i\}L_V(t_i)]^{1-\delta_i}
\end{aligned} \tag{4.19}$$

Considerando covariáveis somente na fração de cura, pode-se reescrever a Equação (4.17) da seginte maneira

$$S_{pop}(t|\mathbf{x}) = p_0(\mathbf{x}) + (1 - p_0(\mathbf{x}))S(t) \tag{4.20}$$

e o modelo de mistura padrão com fragilidade aditiva na presença de covariáveis dado em (4.18) pode ser reeescrito como

$$S_{pop}(t|\mathbf{x}) = p_0(\mathbf{x}) + (1 - p_0(\mathbf{x}))\exp\{-\Lambda_0(t)\}L_V(t). \tag{4.21}$$

Assim, a função de verossimilhança baseada nos dados observados supondo censura

não informativa é dada por

$$\begin{aligned} L(\theta|\mathcal{D}) &= \prod_{i=1}^{n}[f_{pop}(t_i|\mathbf{x}_i)]^{\delta_i}[S_{pop}(t_i|\mathbf{x}_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n}[(1-p_0(\mathbf{x}_i))\lambda(t_i)S(t_i)]^{\delta_i} \times [p_0(\mathbf{x}_i)+(1-p_0(\mathbf{x}_i))S(t_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n}\left[(1-p_0(\mathbf{x}_i))\left(\lambda_0(t_i)-\frac{L_V'(t_i)}{L_V(t_i)}\right)\exp\{-\Lambda_0(t_i)\}L_V(t_i)\right]^{\delta_i} \\ &\quad \times [p_0(\mathbf{x}_i)+(1-p_0(\mathbf{x}_i))\exp\{-\Lambda_0(t_i)\}L_V(t_i)]^{1-\delta_i}. \end{aligned}$$

Para completar a abordagem paramétrica é necessário supor uma distribuição para a fragilidade e para o risco de base.

4.3 Aplicação a dados reais

Os dados dessa aplicação são de um estudo em mulheres diagnosticadas com câncer de mama triplo negativo, que receberam o tratamento de quimioterapia neoadjuvante de 2001 a 2013, no Centro A. C. Camargo, São Paulo, Brasil. Nesse estudo o evento de interesse era a morte devido ao câncer de mama, o tempo de vida considerado foi o decorrido entre a data do diagnóstico até o óbito e os tempos de censura foram definidos como o decorrido entre a data do diagnóstico até o fim do estudo para as pacientes que não apresentaram o evento de interesse, ou da data do diagnóstico até à morte por outra causa.

O estudo contou com 78 pacientes, dentre as quais 32,1% morreram devido ao câncer de mama e 67,9% não apresentaram o evento de interesse. Dentre as covariáveis observadas no estudo utilizou-se nessa aplicação a covariável que capta as características dos lifonodos das cadeias de drenagem linfática do órgão em que o tumor está, denotada por N, em que N0 representa que os linfonos próximos não contêm câncer e N1, N2 e N3 representam o grau de espalhamento do câncer para os linfonodos, de forma que N0 significa que não houve espalhamento do câncer para os linfonodos e N1, N2 e N3 significa que houve espalhamento. Sendo assim categorizou-se N=0 quando não houve espalhamento (N0) e N=1 caso contrário (N1, N2 e N3).

A Tabela 4.1 apresenta a variável utilizada nessa aplicação.

O desfecho final de interesse, denominado de falha, correspondeu ao óbito. Foram

Tabela 4.1: Descrição dos dados

Variável	Descrição	Categoria
Localização do Tumor (N).	N0 ,	0
	N1 ou N2 ou N3	1

censurados os casos em que não foi possível identificar a condição vital, ou seja, não foram observados até a ocorrência da falha. A Tabela 4.2 mostra que do total de 78 pacientes, 25 vieram a óbito e 53 (67,9%) foram censurados, isto é não foram a óbito.

Tabela 4.2: Frequência absoluta e relativa do total de mulheres com câncer de mama

	Status		
No. total de casos	Óbito	Não Óbito	Porcentagem
78	25	53	67,9%

A Tabela 4.3 mostra que mulheres diagnosticadas com câncer de mama e os linfonodos vizinhos não contêm câncer (N0), corresponde a 26,9% e destas 12% foram a óbito. Enquanto que mulheres diagnosticadas com câncer de mama e os linfonodos vizinhos contêm câncer corresponde a 73,1%, e destas 88,0% foram a óbito.

Tabela 4.3: Frequência absoluta e relativa da característica localização do tumor (N)

Característica		Status		Total
		Não óbito	Óbito	
Localização do Tumor (N)				
N0	contagem em	18	3	21
	% Status	34,0%	12,0%	26,9%
N1-N3	contagem em	35	22	57
	% Status	66,0%	88,0%	73,1%
Total	contagem em	53	25	78
	% Status	100,0%	100,0%	100,0%

Para as 21 pacientes na categoria N0 apenas 12,3% vieram à óbito, enquanto que para a outra categoria (N1, N2 e N3) 88%. A Figura 4.3 mostra as curvas de Kaplan-Meier para cada categoria da covariável a respeito dos linfonodos das cadeias de drenagem linfática do órgão. É possível identificar que há uma diferença grande entre sobrevivências estimadas para cada categoria da covariável. Nota-se que a categoria que engloba a localização do tumor N1, N2 e N3 possui a menor expectativa de sobrevivência, provavelmente por se tratar de um tumor que também está localizado nos linfonodos de regiões próximas a mama.

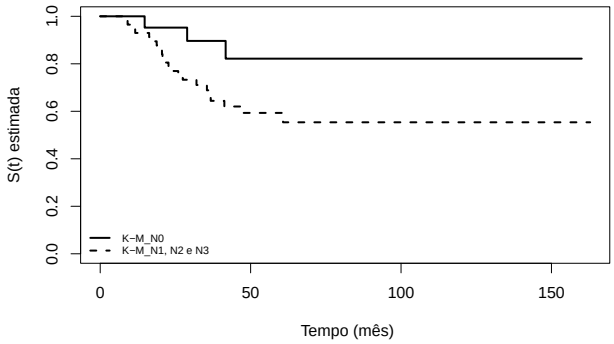

Figura 4.3: Curvas de sobrevivência estimadas por Kaplan-Meier.

4.3.1 Estimação do modelo na presença da covariável "Localização do Tumor"(N).

Para estimar o modelo proposto neste trabalho, modelo de mistura padrão com fragilidade aditiva gama, foram consideradas duas distribuições de risco base: Exponencial e Weibull.

A Tabela 4.4 mostra as estimativas de máxima verossimilhança para cada parâmetro do modelo. É importante ressaltar que o valor estimado da variância $1/\hat{\alpha} = 1/0.002 = 500$ representa a heterogeneidade não observada, ou seja, há uma alta heterogeneidade dos indivíduos e evidências de que existem outros fatores não observados que tenham efeito significativo no tempo de vida das pacientes.

Tabela 4.4: Estimativa de máxima verossimilhança (EMV) e erro padrão (EP) para os parâmetros do modelo (4.21)

	Exponencial		Weibull	
Parâmetros	EMV	EP	EMV	EP
α	0,0020	0,0076	0,0021	0,0085
θ	0,0261	0,0089	0,0329	0,0044
γ	-	-	1,7629	0,3037
β_0	1,2871	0,6498	1,7728	0,6781
β_1	$-1,2871$	0,7218	$-1,3066$	0,7391
AIC	303,3553		293,8634	
BIC	312,7821		305,6469	

Observa-se também que os critérios AIC e BIC são menores para o modelo que utiliza o risco de base da distribuição Weibull, indicando que, por esses critérios, o modelo de mistura padrão com fragilidade aditiva gama e risco de base da distribuição Weibull é melhor do que o modelo com risco de base da distribuição Exponencial.

A Figura 4.4 mostra a curva de sobrevivência estimada pelo modelo de mistura padão com fragilidade Gama com risco base Exponencial e Weibull, respectivamente. Nota-se que há uma certa diferença entre as curvas estimadas pelos dois modelos, observa-se que para o modelo com risco de base Exponencial as curvas estimam valores para a fração de cura abaixo do que a curva de Kaplan-Meier indica. E o comportamento contrário pode ser observado para o modelo com risco de base Weibull. Porém, observa-se que esse último tem curvas que se comportam de forma mais semelhante ao Kaplan-Meier.

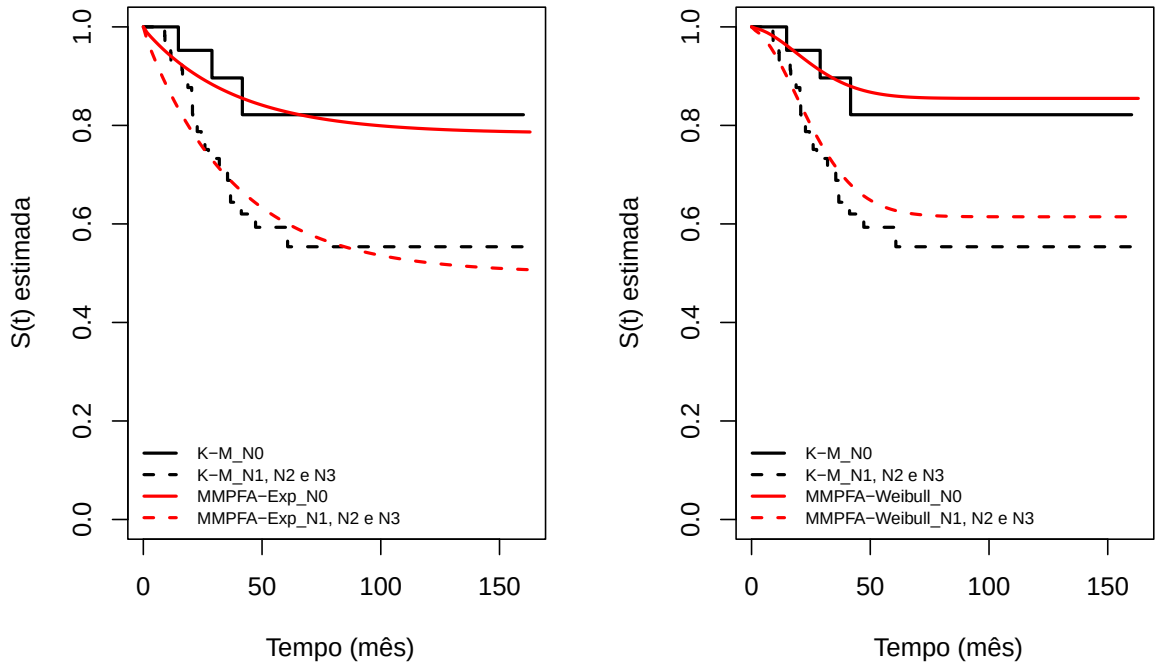

Figura 4.4: Curvas de sobrevivência estimadas por Kaplan-Meier (linhas em preto) e pelo modelo (4.21) (linhas vermelhas) com funções de risco de base das distribuições Exponencial(esquerda) e Weibull (direita).

A partir da função de ligação logística dada em (4.8) obtiveram-se as estimativas para as frações de cura de cada categoria da covariável N na Tabela 4.5, $p_0 = 0.7837$

representando a fração de cura para a mulheres com a localização do tumor $N0$ e $p_1 = 0.500$ representando a fração de cura para as mulheres com a localização do tumor N1-N3. Pode-se observar que a fração de cura para o grupo de mulheres com a localização do tumor $N0$ é maior. Nota-se também que os valores de fração de cura são menores para o modelo com risco de base Exponencial quando comparado ao modelo com Weibull, conforme já mencionado anteriormente. Observa-se que os valores são extremamente próximos e condizentes com as proporções de cura que podem ser observadas na Figura 4.3.

Tabela 4.5: Estimativa de máxima verossimilhança (EMV) para a proporção de cura do modelo (4.21).

	Exponencial		Weibull	
Parâmetros	EMV	EP	EMV	EP
p_0	0,7837	0,1102	0,8548	0,0842
p_1	0,5000	0,1036	0,6145	0,0724

Sendo assim, não há muitos indícios de qual modelo seja o mais adequado para esses dados com a presença da covariável N, porém, por ser um modelo que apresentou menores valores para os critérios AIC e BIC e melhor ajuste das curvas quando comparado ao Kaplan-Meier, decidiu-se que o melhor é o modelo de mistura padrão com fragilidade gama e risco de base Weibull.

4.4 Conclusão

Este capítulo teve o intuito de apresentar o conceito a respeito da metodologia por trás do modelo de mistura padrão com fragilidade aditiva, para isso foram utilizados os conceitos de modelo de fragilidade, explicados no capítulo anterior, e de modelo de longa duração, este que se caracteriza por incorporar indivíduos que não são suscetíveis ao evento de interesse. Além disso, foi realizada uma aplicação do modelo de mistura padrão com fragilidade aditiva em dados reais de mulheres diagnosticadas com câncer de mama triplo negativo, na qual foram consideradas duas distribuições de risco base: Exponencial e Weibull.

Capítulo 5

Conclusão

Este trabalho se propôs a estudar uma modelagem denominado como modelo de mistura padrão com fragilidade aditiva. Essa metodologia engloba duas outras que são muito utilizadas nos estudos de análse de sobrevivência: modelos de longa duração e modelos de fragilidade.

Os modelos de longa duração se diferenciam dos modelos usuais de análise de sobrevivência, pois consideram que nem todos os indivíduos observados no estudo sofreram do evento de interesse, mesmo sendo observados por um longo período de tempo. Sendo assim, esses modelos se caracterizam por incluir em sua modelagem a existência de uma fração de cura, que diz respeito a uma parcela da população que não é suscetível ao evento de interesse esperado, esses indivíduos são considerados como curados. A vantagem de utilizar essa metodologia é o fato dela incorporar a heterogeneidade entre duas subpopulações, curados e não curados. O modelo de mistura padrão é o mais comum para explicar dados de longa duração e utiliza da mistura de distribuições paramétricas, sendo uma função de sobrevivência imprópria para a população total e uma própria para a população de não curados.

Um dos problemas dos modelos de longa duração é o fato de considerarem que todos os indivíduos que sofreram o evento de interesse, não curados, pertencem a uma população homogêna. Porém, há um certo grau de heterogeneidade induzido por fatores de risco que não foram observados. Essa heterogeneidade é incorporada nos modelos de fragilidade que são caracterizados por utilizarem uma variável aleatória que foi observada, que representa informações que não puderam ou que não foram observadas como fatores ambientais, genéticos, informações que por algum motivo não foram incluídas no planejamento do estudo. Este modelo engloba duas fontes de variações: a heterogeneidade entre

os indivíduos causada por covariáveis que não foram observadas e aquela proveniente das covariáveis comuns a indivíudos de um mesmo grupo. Essa fragilidade é introduzida na função de risco e pode ser incluída de forma multiplicativa ou aditiva.

Nos modelos de fragilidade aditiva é observado que quando o tempo do estudo é muito longo a função de sobrevivência tende a zero, sendo assim, não capta a fração de cura daquela população. Com isso, decidiu-se estudar de forma conjunta os modelos de fragilidade aditiva com os modelos de longa duração, utilizando o modelo de mistura padrão. Esse tipo de modelo têm duas características principais que os diferenciam de modelos usuais de sobrevivência: a incorporação de uma parcela dos indivíduos que não apresentam o evento de interesse, mesmo após um longo tempo de acompanhamento, e também a possibilidade de considerar heterogeneidade não observada na modelagem por meio de fatores de riscos não observáveis. Além disso, esse modelo é considerado como uma alternativa ao modelo de riscos proporcionais de Cox, pois não impõe a condição de propocionalidade.

Para verificar a aplicabilidade do modelo de mistura padrão com fragilidade aditiva foi considerada uma aplicação a dados de mulheres diagnosticas com câncer de mama triplo negativo, que receberam quimioterapia neoadjuvante por 12 anos, e o evento de interesse era a morte devido ao câncer de mama. Foi utilizada uma covariável a respeito do espalhamento do câncer e observou-se que aquelas mulheres que tiveram o espalhamento do câncer para outras regiões sofreram mais do evento de interesse, pois sua proporção de cura foi menor do que para aquelas que não tiveram o espalhamento. Aplicou-se a esses dados a metodologia mencionada anteriormente, utilizando a distribuição Gama para a fragilidade e duas distribuições para o risco base: Exponencial e Weibull. Notou-se que os dados possuem uma alta heterogeneidade não observada e que o modelo com risco de base Weibull aparenta ser o mais adequado, visto que teve o menor critérios AIC e BIC. Também na comparação com a curva de Kaplan Maier o modelo considerando o risco de base Weilbul teve um melhor ajuste.

Referências Bibliográficas

Aalen, O. (1978). A model for nonparametric regression analalysis of counting processes. *The Annals of Statistics*, **JSTOR**, 701–726.

Aalen, O. (1980a). A model for nonparametric regression analalysis of counting processes. *Lecture Notes in Statistics*, (2), 1–25.

Aalen, O. (1980b). A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, pages 1–25. Springer.

Berkson, J. e Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(259), 501–515.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, **11**(1), 15–53.

Chen, M.-H., Ibrahim, J. G. e Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**(447), 909–919.

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**(1), 141–151.

Colosimo, E. A. e Giolo, S. R. (2006). *Análise de sobrevivência aplicada*. Editora Blucher.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**(2), 269–276.

Elbers, C. e Ridder, G. (1982). True and spurious duration dependence: The identifiability of the proportional hazard model. *The Review of Economic Studies*, **49**(3), 403–409.

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

Feigl, P. e Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, pages 826–838.

Geraldo Filho, B. (2000). Bagliolo patologia.

Gonçalves, A. T. C., Jobim, P. F. C., Vanacor, R., Nunes, L. N., Albuquerque, I. M. d. e Bozzetti, M. C. (2007). Câncer de mama: mortalidade crescente na região sul do brasil entre 1980 e 2002. *Cadernos de saúde pública. Rio de Janeiro. Vol. 23, n. 8 (ago. 2007), p. 1785-1790*.

Gonzales, J. F. B., Tomazella, V. L. D. e Taconelli, J. P. (2013). Estimação paramétrica do modelo de mistura com fragilidade gama na presença de covariáveis. *Rev. Bras. Biom*, **31**(2), 233–247.

Hougaard, P. (1991). Modelling heterogeneity in survival data. *Journal of Applied Probability*, pages 695–701.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime data analysis*, **1**(3), 255–273.

IMAMA (Porto Alegre, 2014). Instituto da mama rs. câncer de mama metastático. Acesso em 08 de Dez de 2016.

Kaplan, E. L. e Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, **53**(282), 457–481.

Laurie, J. A., Moertel, C. G., Fleming, T. R., Wieand, H. S., Leigh, J. E., Rubin, J., McCormack, G. W., Gerstner, J. B., Krook, J. E. e Malliard, J. (1989). Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. the north central cancer treatment group and the mayo clinic. *Journal of Clinical Oncology*, **7**(10), 1447–1456.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons.

Lin, D. Y. e Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**(1), 61–71.

Lomax, K. S. (1954). Business failures: Another example of the analysis of failure data. *Journal of the American Statistical Association*, **49**(268), 847–852.

Mantel, N. e Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, **22**(4), 719–748.

Milani, E., Calsavara, V. F., Scudilio, J., Tomazella, C. D., Bonini, C. R., Antunes, C. C., Santos, D. G., Zanotti, M. A. e Tomazella, V. (2021). Análise da sobrevida de mulheres diagnosticadas com câncer de mama triplo negativo utilizando modelo de longa duração. *NEXUS Mathematicæ*, **4**.

Ministério da Saúde, I. N. d. C. J. A. G. d. S. (2016). *Estimativa—2016 - Incidência de Câncer no Brasil*.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**(4), 945–966.

Peng, Y. e Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, **56**(1), 237–243.

Rocha, C. M. T. S. (1995). *Modelos com fragilidade em análise de sobrevivência*.

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

Tomazella, V., Louzada-Neto, F. e Silva, G. (2006). Bayesian modeling of recurrent events data with an additive gamma frailty distribution and a homogeneous poisson process. *Journal of Statistical Theory and Applications*, **5**, 417–429.

Tomazella, V. L. D. (2003). *Modelagem de dados de eventos recorrentes via processo de Poisson com termo de fragilidade*. Ph.D. thesis, Instituto de Ciências Matemáticas e de Computação.

Tsodikov, A. D., Yakovlev, A. Y. e Asselain, B. (1996). *Stochastic models of tumor latency and their biostatistical applications*, volume 1. World Scientific.

Vaupel, J. W., Manton, K. G. e Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**(3), 439–454.

Weibull, W. (1939). A statistical theory of the strength of materials, 1939. *Generalstabens Litografiska Anstalts Förlag*.

Weibull, W. (1951). A statistical distribution function of wide applicability. journal of applied mechanics 18: 293-297. *Statistical and Computational Analysis*, **291**.

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