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

Modelo de Mistura Padrão com Fragilidade Gama: Aplicação a Dados de Melanoma

Pamela Cristina Peruchi

Trabalho de Conclusão de Curso

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

Modelo de Mistura Padrão com Fragilidade Gama: Aplicação a Dados de Melanoma

Pamela Cristina Peruchi
Orientadora: Profª Drª Vera L. D. Tomazella

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

São Carlos
Fevereiro de 2024

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Standard Mixture Model with Gamma Frailty: Application to Melanoma Data

Pamela Cristina Peruchi
Advisor: Dr. Prof. Vera L. D. Tomazella

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs - UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
Fevereiro 2024

Pamela Cristina Peruchi

Modelo de Mistura Padrão com Fragilidade Gama: Aplicação a Dados de Melanoma

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Pamela Cristina Peruchi e aprovado pela banca examinadora.

Aprovado em 26 de Janeiro de 2024.

Banca Examinadora:

- Profª Drª Vera L. D. Tomazzela
- Me. Felipe Rodrigues da Silva
- Profº Drº José Carlos Fogo

Agradecimentos

Gostaria de expressar minha sincera gratidão a todas as pessoas que tornaram possível a conclusão deste trabalho acadêmico, principalmente à minha família, que sempre foi meu alicerce. Cada conquista é resultado do apoio que recebo de vocês.

Agradeço aos amigos que estiveram presentes durante essa jornada, pois tive apoio moral e incentivo constante. Dedico especialmente a Alícia Scordamaia por estar comigo desde o começo em cada fase, sendo boa ou ruim,

Obrigado a todos que, de alguma forma, contribuíram para esse marco, me sinto grata por compartilhar este momento com pessoas tão especiais.

Resumo

O câncer é uma doença maligna que pode afetar diversos órgãos. Quando se espalha pelo corpo e atinge outros órgãos, é chamado de metástase, e essa é a principal causa de morte relacionada ao câncer. Embora o câncer de pele seja o mais frequente no Brasil, e corresponda a cerca de 30% de todos os tumores malignos registrados no país, o melanoma representa apenas 4% das neoplasias malignas, de acordo com o Instituto Nacional de Câncer (INCA). A exposição intensa e súbita de radiação ultravioleta (UV) pode provocar queimaduras solares, essa exposição resulta em uma mutação no DNA das células responsáveis pela produção de melanina, desenvolvendo células anormais e levando por sua vez ao desenvolvimento do melanoma. Devido a alta incidência de óbito pelo câncer, é crucial investir em pesquisas, educação e recursos para a prevenção, diagnóstico e tratamento eficaz dessa doença. Dentre as técnicas estatísticas utilizadas para a construção dos modelos, destaca-se a análise de sobrevivência. A aplicação desta técnica na área médica busca estimar, por exemplo, o tempo de sobrevida dos pacientes após o diagnóstico da doença, além de fornecer informações sobre a progressão da doença, a eficácia dos tratamentos e fatores de risco associados. Com base nos dados de sobrevida dos pacientes, é possível identificar padrões e determinar quais características estão associadas a melhores ou piores prognósticos. Neste contexto, este trabalho tem por objetivo considerar um modelo de fração de cura com fragilidade para análise de dados de melanoma. Esse modelo leva em consideração a possibilidade de que alguns pacientes possam ser curados da doença, enquanto outros não. Além disso, incorpora a ideia de fragilidade, que permite levar em conta a variabilidade não capturada pelas variáveis explicativas observadas no estudo. A metodologia proposta será aplicada a uma base de dados da Fundação Oncocentro de São Paulo (FOSP). Essa base de dados contém informações relevantes sobre pacientes com melanoma, incluindo seus tempos de sobrevida e características clínicas.

Palavras-chave: *análise de sobrevivência, melanoma, câncer, fração de cura, fragilidade, modelo de longa duração.*

Abstract

Cancer is a malignant disease that can affect various organs. When it spreads throughout the body and reaches other organs, it is called metastasis, and this is the main cause of cancer-related deaths. Although skin cancer is the most frequent in Brazil, accounting for about 30% of all malignant tumors recorded in the country, melanoma represents only 4% of malignant neoplasms, according to the National Cancer Institute (INCA). Intense and sudden exposure to ultraviolet (UV) radiation can cause sunburns, leading to a mutation in the DNA of melanin-producing cells, resulting in abnormal cells and ultimately leading to the development of melanoma. Due to the high incidence of cancer-related deaths, it is crucial to invest in research, education, and resources for the prevention, diagnosis, and effective treatment of this disease. Among the statistical techniques used for model construction, survival analysis stands out. Applying this technique in the medical field aims to estimate, for example, the survival time of patients after the disease diagnosis, providing information on disease progression, treatment efficacy, and associated risk factors. Based on patient survival data, it is possible to identify patterns and determine which characteristics are associated with better or worse prognoses. In this context, the objective of this work is to consider a cure fraction model with frailty for the analysis of melanoma data. This model takes into account the possibility that some patients may be cured of the disease while others are not. Additionally, it incorporates the idea of frailty, allowing for the consideration of variability not captured by observed explanatory variables in the study. The proposed methodology will be applied to a database from the Oncocentro Foundation of São Paulo (FOSP), containing relevant information about melanoma patients, including their survival times and clinical characteristics.

Keywords: *survival analysis, melanoma, cancer, cure fraction, frailty, long-duration model.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Sendo um dos grandes males da nossa era, o câncer afeta milhões de pessoas no mundo todo e tem impacto na sociedade. Existe uma grande variedade de tipos de câncer, sendo que os fatores ambientais, hereditariedade genética e estilo de vida figuram como suas principiais causas (Anand *et al.*, 2008). No Brasil, dentre os milhares de novos casos de câncer diagnosticados a cada ano, os tipos de cânceres mais comuns são: de pele, de mama, de próstata, de cólon e reto, de testículo, de pulmão e de estômago; somando mais de 65% dos casos, segundo o Instituto Nacional de Câncer (INCA).

O Instituto Nacional de Câncer (INCA) estima que para cada ano do triênio 2020/2022, foram diagnosticados no Brasil 8.450 novos casos de câncer de pele tipo melanoma (4.200 em homens e 4.250 em mulheres). Esses valores correspondem a um risco estimado de 4 casos novos a cada 100 mil pessoas.

O melanoma é 20 vezes mais frequente em pessoas de raça branca do que em pessoas negras. Em geral, o risco de melanoma é de cerca de 2,6% em brancos e 0,1% em negros, além disso é mais comum entre os homens, mas em mulheres com menos de 50 anos de idade as taxas são mais altas. O risco de melanoma aumenta com a idade, sendo a média atual do diagnóstico aos 65 anos, mas o melanoma não é raro entre as pessoas com menos de 30 anos. Na verdade, é um dos cânceres mais frequentes em adultos jovens, especialmente entre as mulheres. (Society, 2023).

Dependendo do tipo de câncer, um número de pacientes pode desenvolver imunidade em relação à doença em estudo ou ser considerado "curado". Nestes casos assume-se que os pacientes não sejam suscetíveis ao evento de interesse (por exemplo, morte ou recidiva da doença). 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, estes são denominados modelos de longa duração ou modelos de fração de cura.

A modelagem da fração de cura, também conhecida como modelagem de longa duração, faz parte da análise de sobrevivência e estuda casos em que, supostamente, existem observações não suscetíveis ao evento de interesse. Boag (1949) foi um dos pioneiros dentro da modelagem de longa duração. Posteriormente, outros modelos foram propostos, como o modelo de mistura padrão por Berkson e Gage (1952); o modelo unificado de fração de cura por Rodrigues *et al.* (2009), dentre outros.

Os modelos convencionais de sobrevivência pressupõem implicitamente que uma população seja homogênea no que diz respeito aos indivíduos suscetíveis ao evento de interesse. No entanto, em muitas situações práticas, as informações dos pacientes podem variar significativamente devido às diferenças em seus estilos de vida, o que resulta em diferentes níveis de risco. Parte dessa variabilidade entre os indivíduos pode ser explicada por variáveis observáveis. No entanto, existe uma parcela de heterogeneidade não observada que pode ser atribuída a diversos fatores, como influências ambientais, genéticas ou informações que não foram levadas em consideração no planejamento. Além disso, em certos cenários, como no caso de modelos de fração de cura, a heterogeneidade na população pode se manifestar de forma distinta, em que uma fração dos pacientes pode ser curada do evento de interesse, enquanto outra fração permanece suscetível.

Portanto, ao considerar a heterogeneidade na análise de sobrevivência, é necessário não apenas examinar as covariáveis observáveis, mas também reconhecer e modelar adequadamente a heterogeneidade não observada, que pode desempenhar um papel fundamental na compreensão das dinâmicas de sobrevivência em determinadas populações. Nestas circunstâncias, é preciso considerar modelos que incorporam heterogeneidade não observável entre os indivíduos, como o modelo de fragilidade (Vaupel *et al.*, 1979).

A fragilidade, nesse contexto, refere-se a uma característica intrínseca de cada unidade do estudo que afeta sua probabilidade de experimentar um evento. Essa característica não é diretamente observável, mas é modelada como um efeito aleatório, que captura a heterogeneidade entre as unidades de estudo. A incorporação da fragilidade no modelo permite levar em conta a variabilidade não observada e não capturada pelas variáveis explicativas, melhorando assim a precisão das estimativas e a capacidade de fazer inferências corretas sobre as associações entre as variáveis explicativas e o evento de interesse (Wienke, 2010).

Em geral, o termo fragilidade é incluído de forma multiplicativa para a função de risco

base e representa uma extensão do modelo de riscos proporcionais introduzido por Cox (1972). De acordo com Wienke (2010), a ideia do modelo de fragilidade é que os indivíduos têm diferentes fragilidades e que o mais frágil falhará mais cedo do que o menos frágil. No estudo de Espirito Santo (2022), a fragilidade discreta é abordada, incorporando uma fração de cura e considerando riscos proporcionais. Essa abordagem analisa a dinâmica do evento de interesse em relação à fragilidade, analisando como a fração de cura e os riscos proporcionais podem influenciar a interpretação das estimativas.

Por outro lado, Taconeli (2013) propõe um modelo de mistura paramétrica que incorpora a fragilidade, levando em consideração covariáveis. Essa abordagem é especialmente relevante quando as características individuais podem influenciar a probabilidade de ocorrência do evento de interesse, além da fragilidade. Outra abordagem significativa é apresentada no estudo de Azevedo Silva (2011), no qual são explorados modelos de fração de cura com fatores latentes competitivos e fragilidade.

É interessante ressaltar que, mesmo ao empregar um único conjunto de dados, é possível explorar uma diversidade de aplicações e abordagens metodológicas. Esse fato é ilustrado de forma vívida nos trabalhos de Calsavara *et al.* (2020) e Molina *et al.* (2021), que demonstram essa versatilidade ao aplicar análises distintas ao mesmo conjunto de dados. Importante notar que, no âmbito desses estudos, a modelagem de fragilidade a longo prazo, utilizando um modelo de riscos não proporcionais, e os modelos de sobrevivência induzidos por fragilidade discreta, com séries de potências modificadas por zero, foram adotados como estratégias analíticas distintas.

Além disso, no contexto deste trabalho, é válido mencionar que o conjunto de dados utilizado por Calsavara *et al.* (2020) e Molina *et al.* (2021) será empregado, considerando um modelo de fração de cura com fragilidade gama, ou seja, apesar da base de dados compartilhada, as análises revelam diferenças significativas, exemplificando a diversidade de interpretações ao explorar abordagens variadas de modelagem em análise de sobrevivência.

1.1 Objetivos

O principal objetivo deste Trabalho de Conclusão de Curso é abordar a metodologia de modelagem de fração de cura com fragilidade e realizar a análise de um conjunto de dados reais de Melanoma, fornecido pela Fundação Oncocentro de São Paulo (FOSP).

O modelo estudado é denominado “Modelo de Mistura Padrão com Fragilidade Gama”, em que será considerada para a modelagem de fração de cura o Modelo de Mistura Padrão proposto por Berkson e Gage (1952) e para a fragilidade a distribuição gama estudado em Wienke (2010).

1.2 Organização do Trabalho

A estrutura deste trabalho é organizada da seguinte maneira. No Capítulo 2, será apresentada uma revisão da literatura com os conceitos básicos da análise de sobrevivência e metodologias de modelagem de longa duração. No Capítulo 3, será introduzido o modelo de fração de cura com fragilidade. No Capítulo 4, será apresentado o conjunto de dados que será analisado no trabalho e no Capítulo 5 apresentamos conclusões e propostas futuras.

Capítulo 2

Metodologia

Neste capítulo, serão apresentados os conceitos básicos da teoria de análise de sobrevivência. Serão abordados os tipos de censura, as funções básicas de sobrevivência, o estimador de *Kaplan-Meier*, o modelo de *Cox*, o modelo de mistura padrão e a fragilidade. Além disso, serão abordadas algumas distribuições relevantes no contexto da análise de sobrevivência.

2.1 Conceitos Básicos de Análise de Sobrevivência

A análise de sobrevivência incorpora técnicas para analisar dados que estão relacionados ao tempo até a ocorrência de um evento de interesse como, por exemplo, o tempo de sobrevivência, vida ou falha. Consiste em uma coleção de procedimentos estatísticos para análise de dados relacionados ao tempo de ocorrência de um determinado evento de interesse a partir de um momento inicial previamente definido. A análise de sobrevivência tem como objetivo principal estudar a função de risco ou de sobrevivência de indivíduos, permitindo comparar as distribuições de sobrevivência entre diferentes grupos de pacientes em um experimento, por exemplo. Além disso, essas aplicações auxiliam na tomada de decisões clínicas e no desenvolvimento de estratégias de prevenção e tratamento nos casos clínicos.

A análise de sobrevivência se tornou proeminente a partir do meio do século XX e é fundamental em diversos campos, como medicina, epidemiologia, engenharia e ciências sociais, contribuindo para a compreensão dos tempos de vida.

Na análise de sobrevivência, os três tipos mais comuns de estudos são os descritivos, com apenas uma amostra; os comparativos, com dois ou mais grupos; e o estudo de

coorte. Em estudos de coorte, um grupo de indivíduos é acompanhado ao longo de um certo tempo, são observados os fatores de interesse e o desenvolvimento de eventos ou desfechos específicos. Esses estudos podem ser prospectivos, nos quais os indivíduos são acompanhados a partir do presente para o futuro, geralmente por meio de coleta de dados em tempo real, como em ensaios clínicos em andamento. Por outro lado, os estudos retrospectivos envolvem a coleta de dados a partir de registros históricos. Um exemplo disso é um estudo que compara o histórico de exposição ao tabagismo entre pacientes diagnosticados com câncer de pulmão (casos) e indivíduos sem câncer de pulmão (controles) com base em registros médicos e questionários preenchidos no passado.

2.1.1 Falha ou Desfecho

Em um estudo de análise de sobrevivência, a falha ou desfecho é o evento de interesse que representa o objetivo principal do estudo. A definição do evento de interesse é essencial para estabelecer critérios específicos que determinem o momento preciso em que esse evento ocorre afim de garantir a interpretação adequada dos resultados. Existem diversos tipos de falhas ou desfechos que podem ser considerados em um estudo de sobrevivência, alguns exemplos comuns incluem o aparecimento de efeitos colaterais de um medicamento, o diagnóstico de uma doença, o surgimento de um tumor, recidiva de uma doença, o óbito do paciente, o nascimento e a cura. A escolha do desfecho dependerá dos objetivos específicos da pesquisa, é comum encontrar eventos censurados em estudos de sobrevivência

2.1.2 Censura

Uma das principais características dos dados de sobrevivência é a presença de censura, que ocorre quando o tempo até o evento de interesse para alguns indivíduos não é observado, isso pode acontecer devido à perda de acompanhamento (*follow-up*), desistência do participante ou remoção do estudo por outras causas. A presença de censura traz desafios à análise, mas também fornece informações valiosas. Isso pode ocorrer em várias situações, como quando o estudo é encerrado antes que todas as falhas tenham ocorrido ou quando há indivíduos vivos cujos tempos de sobrevivência não foram completamente observados. Sem a censura, as técnicas estatísticas clássicas, como regressão e planejamento de experimentos, poderiam ser aplicadas diretamente aos dados. No entanto, a presença

da censura requer métodos específicos de estimação, como o estimador de *Kaplan-Meier* e a regressão de *Cox*.

Existem diferentes tipos de censura, Klein e Moeschberger (2003) as classificaram da seguinte forma:

- **Censura Tipo I:** A censura do Tipo I ocorre quando o evento é observado apenas se ocorrer antes de um tempo pré-especificado. Não há informações sobre o momento exato de um evento para alguns indivíduos, já que a observação deles termina antes que esse evento ocorra. Por exemplo, em um estudo de sobrevivência de pacientes, se alguns participantes permanecem vivos até o fim do estudo, o tempo de sobrevivência deles é considerado censurado, indicando que a observação foi encerrada sem registrar o evento de interesse.

- **Censura Tipo II:** A censura do Tipo II ocorre quando o evento de interesse acontece antes do início do estudo e o tempo desse evento não é observado. O estudo continua até que falhem os primeiros r indivíduos (em que r é um número pré-definido). Experimentos com censura do Tipo II são comuns em testes de vida de equipamentos, economizando tempo ao encerrar o teste quando r dos itens totais falham. Ao contrário da censura do Tipo I, em que o tempo de censura é fixo, no Tipo II, o número de falhas e de observações censuradas são constantes, enquanto o tempo de censura é aleatório;

- **Censura Aleatória:** A censura aleatória é um tipo de censura em que a probabilidade de um evento ser censurado não está relacionada com o tempo, ou seja, a censura acontece de forma aleatória e independente do decorrer do tempo ou da ocorrência do evento de interesse. Por exemplo, quando a perda de acompanhamento ocorre de forma imprevisível e independente do tempo.

Matematicamente, a censura é representada através de 2.1:

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é tempo de falha, } T_i \leq C_i \\ 0, & \text{se } t_i \text{ é tempo de censura, } T_i > C_i, \end{cases} \tag{2.1}$$

em que as variáveis T_i e C_i representam, respectivamente, o tempo até o desfecho do paciente i, sendo $i = 1, ..., n$, e o tempo de censura desse mesmo paciente, ambas são

variáveis aleatórias. O tempo t_i observado é o menor valor entre o tempo de desfecho e o tempo de censura para cada paciente, ou seja, $t_i = \min[T_i, C_i]$.

2.1.3 Funções Básicas

As funções básicas da análise de sobrevivência são fundamentais para compreender eventos ao longo do tempo e analisar dados censurados.

A **função de sobrevivência** descreve a probabilidade de um evento não ter ocorrido até um determinado tempo, ou seja, a probabilidade do indivíduo sobreviver por um período superior a t, isto é,

$$\begin{aligned} S(t) = P(T > t) &= \int_t^{\infty} f(u)du \\ &= 1 - F(t), \end{aligned} \tag{2.2}$$

sendo $F(t)$ a função de distribuição acumulada da variável aleatória T,

$$F(t) = P(T \leq t) = \int_0^t f(u)du. \tag{2.3}$$

A função de sobrevivência apresenta as seguintes propriedades:

1. $S(0) = 1$;
2. $S(t)$ não é crescente;
3. $\lim_{t\to\infty} S(t) = 0$.

A Figura 2.1 mostra o comportamento da curva de sobrevivência ao longo do tempo, nesse caso a função de sobrevivência chega ao valor 0.

A **função de risco** ou função de taxa de falha descreve a taxa instantânea na qual o evento de interesse (falha, morte, ou outro desfecho) ocorre em um determinado ponto no tempo, dado que o indivíduo tenha sobrevivido até esse momento, ou seja, é a probabilidade instantânea de ocorrer o evento considerando que o indivíduo tenha sobrevivido até o instante de tempo analisado:

$$h(t) = \lim_{\Delta(t)\to 0} \frac{P(t \leq T < t + \Delta(t) | T > t)}{\Delta(t)} = \frac{f(t)}{S(t)}. \tag{2.4}$$

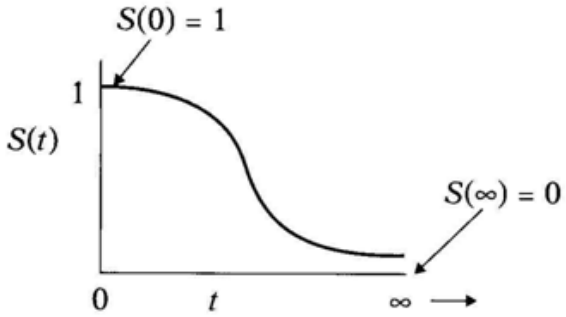

Figura 2.1: Exemplo de curva de sobrevivência.

A função de risco pode variar em diferentes comportamentos, como crescente, decrescente, constante e em formas não monótonas, como a "curva da banheira". Essa curva exemplifica a função de risco de mortalidade em seres humanos que, por exemplo, possui três fases (nascimento, fase adulta e envelhecimento) em que tem alta taxa de falha no início, declínio e estabilidade em seguida, e posteriormente um aumento na taxa de falha, conforme a Figura 2.2.

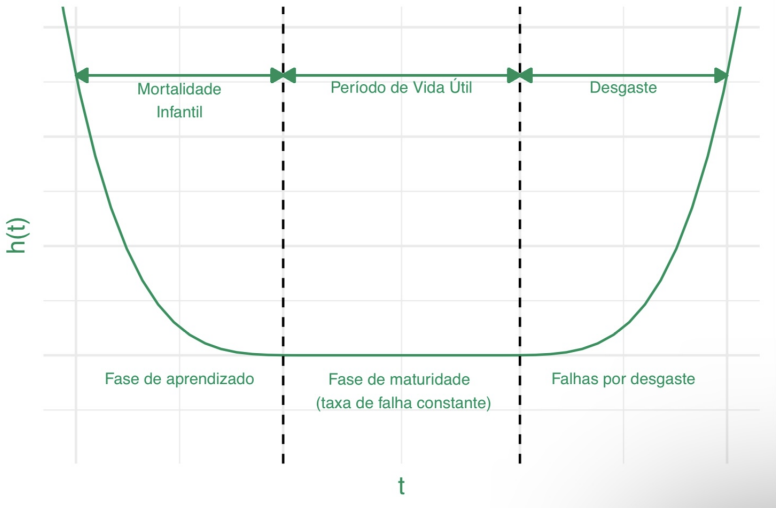

Figura 2.2: Curva de taxa de falha não monótona que descreve a taxa de mortalidade humana. Fonte: Slides da aula de análise de sobrevivência (Vera Tomazella, 2022).

A **função de risco acumulado** representa a probabilidade acumulada da ocorrência de um evento de interesse até um determinado momento no tempo, considerando a sobrevivência dos indivíduos até esse ponto. Isto é,

$$H(t) = \int_0^t h(u)du = -\log(S(t)). \tag{2.5}$$

A função de risco acumulado é importante em análises gráficas para verificar a ade-

quação de modelos estatísticos.

Algumas relações matemáticas importantes entre as equações definidas anteriormente:

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

$$S(t) = \exp\{-H(t)\}, \tag{2.7}$$

e

$$H(t) = -log(S(t)). \tag{2.8}$$

2.1.4 Estimador de *Kaplan-Meier*

O estimador de *Kaplan-Meier* é uma técnica não paramétrica utilizada na análise de sobrevivência, com ele é possível estimar a função de sobrevivência em estudos de acompanhamento ao longo do tempo que contém censura. Esse estimador, proposto por Kaplan e Meier (1958) para estimar a função de sobrevivência, é especialmente útil quando os dados apresentam observações incompletas, geralmente aplicado em estudos médicos, ensaios clínicos e pesquisas que envolvem o tempo até ocorrência de eventos, como mortalidade, recidiva de doenças e falhas em equipamentos.

Suponha que existem n itens sob teste e k $(\leq n)$ falhas distintas nos tempos $t_1 \leq t_2 \leq ... \leq t_k$. Ocasionalmente, pode ocorrer mais de uma falha simultaneamente, denotando de empate. Dessa forma, o estimador de *Kaplan-Meier*, também conhecido como estimador de produto-limite, é definido por:

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

em que d_j denota o número de falhas no tempo t_j e n_j denota o número de itens sob risco, que não falhou e não foi censurado até o momento imediatamente anterior em t_j, $j = 1, .., k$. A equação 2.9 descreve uma função em formato de escada, representando os tempos de falha observados.

O estimador de *Kaplan-Meier* possui as seguintes propriedades:

1. Não-viesado: A estimativa da função de sobrevivência não apresenta um erro sistemático;

2. Fracamente consistente: O estimador é consistentemente próximo do valor verda-

deiro da função de sobrevivência à medida que o tamanho da amostra aumenta, ou seja, a medida que mais dados são incluídos na análise, a estimativa da função de sobrevivência tende a se aproximar da verdadeira função de sobrevivência da população;

3. Possui distribuição assintótica Normal: Quando o tamanho da amostra é grande, a distribuição do estimador de *Kaplan-Meier* se aproxima de uma distribuição Normal;

4. O estimador de *Kaplan-Meier* é a estimativa de máxima verossimilhança da função de sobrevivência $S(t)$ com base nos dados de tempo de sobrevivência e censura.

2.1.5 Estimação de Máxima Verossimilhança

A estimação de parâmetros é o método utilizado para determinar características na amostra que são consideradas representativas ou próximas dos valores reais que existem na população. Isso nos ajuda a fazer inferências sobre a totalidade da população com base nos dados disponíveis em uma parte dela. Existem vários métodos de estimação utilizados para obter essas estimativas e um deles é a estimação de máxima verossimilhança (EMV).

Na análise de sobrevivência, o objetivo é estimar a função de sobrevivência (2.2), que descreve probabilidade de um evento não ter ocorrido até um determinado tempo. Para isso, supõe-se que o tempo até o evento, representado pela variável aleatória T, segue uma distribuição desconhecida com parâmetros a serem estimados. O método de máxima verossimilhança é utilizado para obter essas estimativas, permitindo lidar com dados censurados e proporcionando resultados para amostras grandes. O estimador de máxima verossimilhança é encontrado ao maximizar a função de verossimilhança. Quando não há censuras nos dados, conforme Bolfarine e Sandoval (2001), a função de verossimilhança é:

$$L(\boldsymbol{\theta}; t) = \prod_{i=1}^{n} f(t_i; \boldsymbol{\theta}), \tag{2.10}$$

em que $f(t; \boldsymbol{\theta})$ é a função densidade de probabilidade e $\boldsymbol{\theta}$ é o vetor de parâmetros.

A função de verossimilhança (2.10) mede a contribuição de cada observação não-censurada através de sua função de densidade. Por outro lado, para as observações censuradas, a contribuição é dada por sua função de sobrevivência $S(t)$, pois elas apenas fornecem a informação de que o tempo de falha é maior que o tempo de censura observado. Pode-se dividir as observações das amostras aleatórias em dois grupos, as censuradas e

não censuradas, sendo a função de verossimilhança com dados censurados dada por

$$L(\boldsymbol{\theta};t) \propto \prod_{i=1}^{n}[f(t_i|\boldsymbol{\theta})]^{\delta_i}[S(t_i|\boldsymbol{\theta})]^{1-\delta_i} = \prod_{i=1}^{n}[h(t_i|\boldsymbol{\theta})]^{\delta_i}S(t_i|\boldsymbol{\theta}), \tag{2.11}$$

em que δ_i é a função indicadora de censura. A expressão (2.11) é válida para os tipos de censura I, II e aleatória.

Os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\theta}$ que maximizam $L(\boldsymbol{\theta})$, porém é equivalente utilizar o logaritmo da função 2.11, pois os valores que maximizam a função de verossimilhança e os valores que maximizam a função de log-verossimilhança são os mesmos, isto é, $l(\boldsymbol{\theta}) = log(L(\boldsymbol{\theta}))$. Os estimadores são encontrados resolvendo o sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial log(L(\boldsymbol{\theta};t))}{\partial \boldsymbol{\theta}} = \frac{\partial l(\boldsymbol{\theta};t)}{\partial \boldsymbol{\theta}}. \tag{2.12}$$

Devido a sua complexidade, o sistema (2.12) não resulta em formas algébricas fechadas na maioria dos casos, portanto é necessário o uso de métodos numéricos para realizar a estimação.

2.2 Modelos Probabilísticos

Embora exista uma série de modelos probabilísticos, alguns deles são destaque por sua comprovada adequação a várias situações práticas. Nesta seção serão apresentadas algumas das principais distribuições de probabilidade utilizadas em análise de sobrevivência.

2.2.1 Distribuição Exponencial

A distribuição exponencial é um dos modelos probabilísticos mais simples utilizados para descrever o tempo de falha. Esse modelo apresenta um único parâmetro e possui a propriedade única de ter a função de risco constante. A função de densidade de probabilidade para a variável aleatória tempo de falha T com distribuição exponencial é dada por:

$$f(t) = \lambda e^{-\lambda t}, \quad t \geq 0 \quad \text{e} \quad \lambda > 0. \tag{2.13}$$

A função de sobrevivência $S(t)$ e função de risco $h(t)$ são, respectivamente,

$$S(t) = e^{-\lambda t} \tag{2.14}$$

e

$$h(t) = \lambda. \tag{2.15}$$

Na Figura 2.3 é possível notar o comportamento das funções, graficamente, quando o valor de λ é alterado.

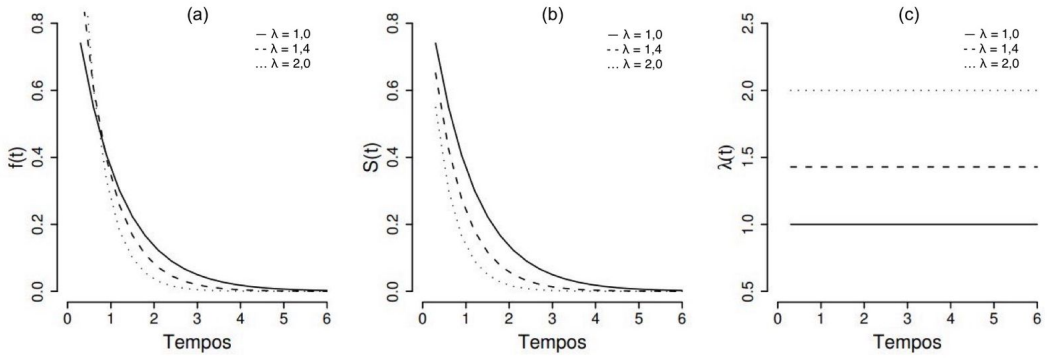

Figura 2.3: Funções de densidade (a), sobrevivência (b) e risco (c) para a distribuição exponencial.

A simplicidade e a propriedade de taxa de falha constante tornam a distribuição exponencial uma escolha conveniente para modelar eventos que ocorrem de forma aleatória ao longo do tempo, como a chegada de clientes, falhas em sistemas, ou qualquer evento que não siga um padrão regular e ocorra em momentos diferentes ao longo do tempo.

2.2.2 Distribuição Gama

A distribuição gama, que inclui a exponencial (2.13) como um caso especial, foi utilizada por Brown e Flood (1947) para descrever o tempo de vida de copos de vidro circulando em uma cafeteria. Essa distribuição tem sido usada em problemas de confiabilidade e sobrevivência, pois se ajusta adequadamente a uma variedade de fenômenos, inclusive em problemas da área médica.

A função densidade da distribuição gama é expressa por:

$$f(t; k, \lambda) = \frac{\lambda^k}{\Gamma(k)} t^{k-1} \exp\{-t\lambda\}, \quad t > 0, \quad k, \lambda > 0. \tag{2.16}$$

em que $k > 0$ é chamado parâmetro de forma, $\lambda > 0$ parâmetro de escala e $\Gamma(k)$ é a

função gama. A função de sobrevivência desta distribuição é dada por:

$$S(t) = \int_t^{\infty} \frac{\lambda^k}{\Gamma(k)} u^{k-1} exp\{-u\lambda\} du. \tag{2.17}$$

Na Figura 2.4 é possível notar o comportamento das funções, graficamente, quando os valores dos parâmetros são alterados.

Segundo Colosimo e Giolo (2006), a função de risco (2.4), apresenta um padrão crescente ou decrescente convergindo para um valor constante quando t cresce de 0 a infinito. Quando $k > 1$, a taxa de falha cresce monotonicamente de 0 até λ, ou seja, a taxa de falha aumenta conforme o passar do tempo. Se k está entre 0 e 1, a taxa de falha decresce monotonicamente de infinito até λ, o que implica que a taxa de falha diminui à medida que o tempo passa. Quando $k = 1$, a taxa de falha é constante, pois tem-se a distribuição exponencial como um caso especial de gama. Se o parâmetro k assume apenas valores inteiros, então se torna uma distribuição Erlang (Lee, 1980).

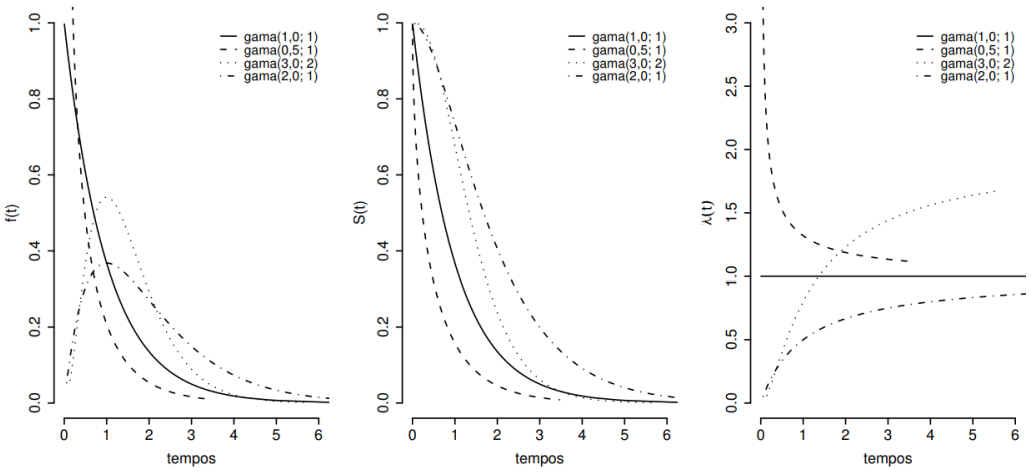

Figura 2.4: Funções de densidade, sobrevivência e risco para a distribuição gama, respectivamente.

2.2.3 Distribuição Gama Generalizada

Essa distribuição, introduzida por Stacy (1962), é caracterizada por três parâmetros, γ, k e λ, todos positivos. Sua função de densidade é dada por:

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

Para a distribuição gama generalizada, λ é um parâmetro de escala, enquanto γ e k

são parâmetros de forma. A partir da função de densidade da gama generalizada, pode-se destacar os seguintes casos:

i) para $\gamma = k = 1$ tem-se $T \sim Exp(\lambda)$;

ii) para $k = 1$ tem-se $T \sim Weibull(\gamma, \lambda)$;

iii) para $\gamma = 1$ tem-se $T \sim Gama(k, \lambda)$.

É possível notar que a distribuição gama generalizada pode se adaptar a diferentes distribuições, dependendo dos valores atribuídos aos parâmetros γ e k.

2.2.4 Distribuição Weibull

A distribuição Weibull (Weibull, 1939) é utilizada em estudos biomédicos e industriais devido à sua versatilidade. Ela se destaca por sua capacidade de modelar comportamentos de falha ao longo do tempo, tendo capacidade de representar taxas de falha que podem variar, seja aumentando, diminuindo ou mantendo-se constantes.

Quando uma variável aleatória T segue uma distribuição Weibull, sua função de densidade de probabilidade pode ser expressa como:

$$f(t) = \nu\lambda(\lambda t)^{\nu-1} \exp\left[-(\lambda t)^{\nu}\right], \quad \nu > 0, \lambda > 0, t > 0, \tag{2.19}$$

sendo ν o parâmetro de forma e λ o parâmetro de escala, ambos com valores positivo. As funções de sobrevivência, de risco e de risco acumulado são, respectivamente,

$$S(t) = \exp[-(\lambda t)^{\nu}], \tag{2.20}$$

$$h(t) = \nu\lambda(\lambda t)^{\nu-1}, \tag{2.21}$$

$$H(t) = (\lambda t)^{\nu}, \tag{2.22}$$

para $t \geq 0$. Quando $\nu = 1$ tem-se a distribuição exponencial, sendo essa um caso particular da distribuição Weibull.

Na Figura 2.5 é possível notar o comportamento das funções, graficamente, quando o valor dos parâmetros é alterado.

Além das distribuições anteriormente descritas, existem outras para modelar o tempo de falha em análise de sobrevivência. Algumas dessas distribuições incluem a normal

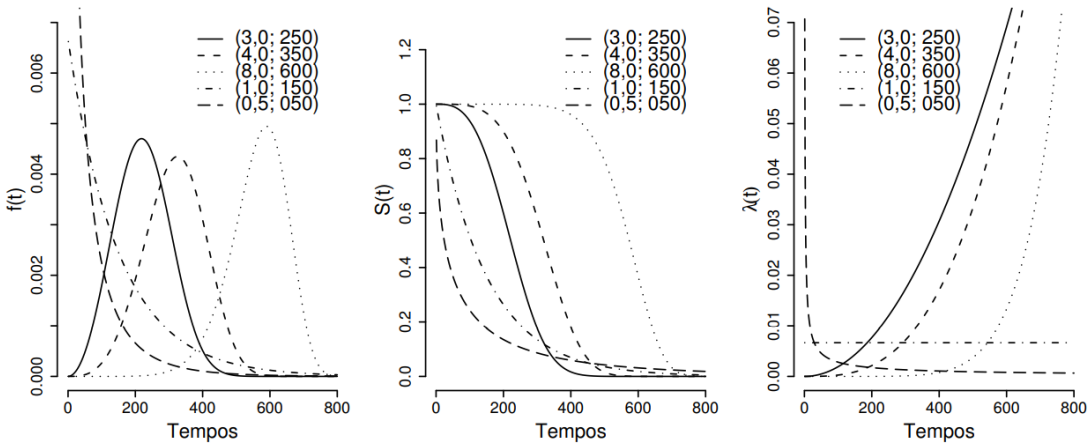

Figura 2.5: Funções de densidade, sobrevivência e risco, respectivamente, para a distribuição Weibull.

inversa, Gompertz, log-gama, entre outras. A escolha da distribuição adequada depende das características específicas dos dados e do fenômeno em estudo. Cada distribuição possui propriedades distintas e pode se ajustar melhor a diferentes tipos de dados.

2.3 Modelo de Riscos Proporcionais de *Cox*

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

$$h(t|\mathbf{x}) = h_0(t)g(\boldsymbol{\beta}'\mathbf{x}), \tag{2.23}$$

em que $h_0(t)$ é a função de risco base, ou seja, é o risco de um indivíduo com covariáveis com $\mathbf{x} = 0$, e $\mathbf{x}$ é o vetor de covariáveis. Já o componente paramétrico da função $(g(\boldsymbol{\beta}'\mathbf{x}))$ é sempre não-negativo e frequentemente utilizado multiplicando a função de risco base, sendo,

$$g(\boldsymbol{\beta}'\mathbf{x}) = \exp\{\boldsymbol{\beta}'\mathbf{x}\} = \exp\{\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k\}, \tag{2.24}$$

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

O modelo de *Cox* (2.23) é dito semi-paramétrico, pois a função de risco de base não é paramétrica, o que significa que sua forma não é fixada com parâmetros específicos definidos previamente. Sendo assim, esse modelo tem mais flexibilidade para capturar padrões complexos de risco ao longo do tempo do que o modelo paramétrico. Ademais, o modelo também é chamado 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,

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'x_i)}{h_0(t)\exp(\boldsymbol{\beta}'x_j)} = \frac{\exp(\boldsymbol{\beta}'x_i)}{\exp(\boldsymbol{\beta}'x_j)} = \exp[\boldsymbol{\beta}'(x_i - x_j)], \tag{2.25}$$

não depende do tempo. Nesse sentido, se um indivíduo no início do estudo possui um risco de morte igual 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 função de risco

As funções relacionadas a $h_0(t)$ referem-se basicamente a:

- Função de risco acumulada de base

$$H_0(t) = \int_0^t h_0(u)du. \tag{2.26}$$

- Função de sobrevivência de base

$$S_0(t) = \exp\{-H_0(t)\}. \tag{2.27}$$

Dessa forma, com covariáveis, temos:

- Função de risco acumulada

$$H(t|\mathbf{x}) = H_0(t)\exp\{\boldsymbol{\beta}'\mathbf{x}\} \tag{2.28}$$

- Função de sobrevivência

$$S(t|\mathbf{x}) = [S_0(t)]^{\exp\{\boldsymbol{\beta}'\mathbf{x}\}} \tag{2.29}$$

2.3.2 Estimação do Modelo de *Cox*

Os coeficientes de regressão $\boldsymbol{\beta}$'s são as quantidades de maior interesse na modelagem 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 para que o modelo seja ajustado. Nesse contexto, é necessário um método de estimação para fazer inferências neste modelo e o método de máxima verossimilhança (2.1.5) é o mais utilizado nessas ocasiões. Contudo, com a presença do componente não-paramétrico $h_0(t)$ na função de verossimilhança, o método torna-se inapropriado para o uso.

Portanto, para solucionar o problema, Cox (1975) propôs uma solução 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 que em uma amostra de n indivíduos existam $k \leq n$ falhas distintas nos tempos $t_1 < t_2 < ... < t_k$. A função de verossimilhança é dada por,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp\{\boldsymbol{\beta}' x_i\}}{\sum_{j \in S(t_i)} \exp\{\boldsymbol{\beta}' x_i\}} \right)^{\delta_i}, \tag{2.30}$$

em que 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 $h_0(t)$ desaparece.

Então, a função de verossimilhança (2.30), 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 são obtidos resolvendo o sistema de equações definido por $U(\boldsymbol{\beta}) = 0$, em que $U(\boldsymbol{\beta})$ é vetor de derivadas de primeira ordem da função $\log(L(\boldsymbol{\beta}))$, isto é,

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{j \in S(t_i)} x_j \exp\{\hat{\boldsymbol{\beta}}' x_j\}}{\sum_{j \in S(t_i)} \exp\{\hat{\boldsymbol{\beta}}' x_j\}} \right] = 0. \tag{2.31}$$

2.3.4 Propriedades dos Estimadores de Verossimilhança para o Modelo de Cox

Cox (1975) e outros autores estudaram as propriedades dos estimadores de máxima verossimilhança parcial e sob condições apropriadas chegaram a algumas conclusões, os estimadores podem ser considerados consistentes e assintoticamente normais. Dessa forma, é possível realizar com mais facilidade testes de hipóteses. Para testar a hipótese, pode-se usar as estatísticas de *Wald*, razão de verossimilhança e escore.

O **teste de *Wald*** é utilizado para testar hipóteses relativas a um único parâmetro, ou seja,

$$H_0 : \beta_j = \beta_{0j}$$

$$W = \frac{(\hat{\beta}_j - \beta_{0j})^2}{\hat{Var}(\hat{\beta})} \sim \chi^2_1, \tag{2.32}$$

em que β_j representa o estimador do parâmetro específico na posição j, β_{0j} é um valor específico (geralmente um valor proposto sob a hipótese nula H_0 para o parâmetro) e os valores de $W > \chi^2_{1,1-\alpha}$ indicam a rejeição de H_0.

2.3.5 Estimativa das funções relacionadas a função de risco

Como $h_0(t)$ não é especificado parametricamente, os estimadores que devem ser utilizados para essas quantidades são estimadores não-paramétricos. Um estimador simples, proposto para a função de risco de base acumulada $H_0(t)$ é o estimador de Breslow, que é uma função escada com saltos nos distintos tempos de falha e é expresso por

$$\hat{H}_0(t) = \sum_{j:t_j<t} \frac{d_j}{\sum_{I \in R_j} \exp\{x'_j \hat{\beta}\}}, \tag{2.33}$$

em que d_j é o número de falhas em t_j, sendo que j varia até o número total de eventos de falha observados. Por consequência, as funções de sobrevivência, $S_0(t)$ e $S(t)$ podem ser, respectivamente, estimadas por

$$\hat{S}_0(t) = \exp\{-\hat{H}_0(t)\}, \tag{2.34}$$

$$\hat{S}(t) = [\hat{S}_0(t)]^{\exp\{x'_l \hat{\beta}\}} \tag{2.35}$$

2.3.6 Interpretação dos Coeficientes

Os coeficientes do vetor $\boldsymbol{\beta}$ no modelo de regressão de *Cox* são indicadores dos efeitos das covariáveis na taxa de falha ao longo do tempo. Cada coeficiente refere-se a uma covariável específica e representa o impacto dessa variável na taxa de falha, podendo acelerá-la ou desacelerá-la.

Quando um coeficiente β possui um valor positivo, indica que a covariável associada contribui para um aumento na taxa de falha, ou seja, está relacionada a um maior risco de ocorrência do evento. Por outro lado, se o coeficiente possui um valor negativo, sugere que a covariável está associada a uma redução na taxa de falha, representando um fator de proteção contra o evento.

Ao calcular a exponencial do coeficiente (e^{β}), tem-se o risco relativo associado à covariável. Se este valor é superior a 1, indica um sobrerisco, significando que a presença da covariável está relacionada a um aumento no risco do evento. Por outro lado, se o valor está entre 0 e 1, sugere uma proteção, indicando que a covariável está associada a uma redução no risco de ocorrência do evento.

2.4 Modelos de *Cox* Paramétricos

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

$$h(t|x_i) = h_0(t)g(\boldsymbol{\beta}' x_i), \quad i = 1, 2, ..., n, \tag{2.36}$$

e como nesse modelo h_0 assume uma forma paramétrica, temos 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 dado por

$$S(t|x_i) = [S_0(t)]^{e^{\boldsymbol{\beta}' x_i}}, \tag{2.37}$$

enquanto a função de verossimilhança é

$$L(\boldsymbol{\beta}, \mathbf{x}) = \prod_{i=1}^{n} [h_0(t) \exp\{\boldsymbol{\beta}'\mathbf{x}\}]^{\delta_i} [S_0(t_i)]^{\exp\{\boldsymbol{\beta}'\mathbf{x}\}}, \tag{2.38}$$

em que $h_0(t)$ é a função de risco de base e $S_0(t)$ é a função de sobrevivência de base.

Nos **modelos de riscos proporcionais exponencial**, têm-se que a função de risco base da distribuição exponencial é

$$h_0(t) = \lambda, \tag{2.39}$$

a função de risco e sobrevivência condicionais do modelo de *Cox* são respectivamente,

$$h(t|x_i) = \lambda \exp(\boldsymbol{\beta}' x_i) \tag{2.40}$$

$$S(t|x_i) = [\exp(-\lambda t)]^{\exp(\boldsymbol{\beta}' x_i)}. \tag{2.41}$$

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

$$L(\boldsymbol{\beta}, x) = \prod_{i=1}^{n} \left[\lambda \exp\left(\boldsymbol{\beta}' x_i\right)\right]^{\delta_i} \left[\exp(-\lambda t)\right]^{\exp(\boldsymbol{\beta}' x_i)}, \tag{2.42}$$

em que δ_i é o indicador de censura.

Nos **modelos de riscos proporcionais Weibull**, têm-se a função de risco base de Weibull,

$$h_0(t) = \nu\lambda(\lambda t)^{\nu-1}, \tag{2.43}$$

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

$$h(t|x_i) = \nu\lambda(\lambda t)^{\nu-1} \exp(\boldsymbol{\beta}' x_i), \tag{2.44}$$

$$S(t|x_i) = [\exp(\nu\lambda(\lambda t))^{\nu}]^{\exp(\boldsymbol{\beta}' x_i)}. \tag{2.45}$$

A função de verossimilhança desse modelo é expressa como

$$L(\boldsymbol{\beta}, \mathbf{x}) = \prod_{i=1}^{n} \left[\nu\lambda(\lambda t)^{\nu-1} \exp(\boldsymbol{\beta}' x_i)\right]^{\delta_i} \left[\exp(\nu\lambda(\lambda t))^{\nu}\right]^{\exp(\boldsymbol{\beta}' x_i)}. \tag{2.46}$$

2.5 Modelos de Longa Duração

O modelo de longa duração fornece maior flexibilidade para capturar padrões complexos de risco ao longo do tempo, tornando-o uma técnica útil para estudos que envolvem eventos de longa duração e comportamentos variáveis da taxa de falha. Ao contrário dos modelos tradicionais, como a distribuição exponencial (2.13) que assume uma taxa de

falha constante, o modelo de longa duração permite que a taxa de falha varie conforme o tempo.

Na análise de sobrevivência, existem casos em que os indivíduos do estudo são considerados imunes ou curados ao evento de interesse (Ibrahim *et al.*, 2014), isso ocorre quando, independentemente do período de tempo do estudo, o evento de interesse não ocorre para parte dos indivíduos. Essa característica é abordada de forma adequada por meio dos modelos de longa duração, especialmente em estudos que envolvem doenças crônicas, em que a taxa de falha pode ser maior em determinados períodos e menor em outros.

Nos modelos de longa duração ou fração de cura, a curva de sobrevivência, estimada pelo método de *Kaplan-Meier*, apresenta um comportamento distinto. Nesse caso, a curva não se estabiliza em zero, mas na fração de cura, ou seja, conforme o tempo aumenta, a estabilidade é contínua, representando a proporção da população considerada imune ou curada ao evento de interesse. Esses modelos podem ser ajustados por vários métodos presentes na literatura, como as abordagens propostas por Lawless (2011) e Chen *et al.* (1999), além de diversas variações e extensões dos modelos de longa duração desenvolvidos.

Nesta seção, será apresentado o modelo de mistura padrão. Esse modelo desempenha um papel importante em situações em que uma parte da população é considerada imune ou curada ao evento de interesse.

2.5.1 Modelo de Mistura Padrão

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

A função de sobrevivência imprópria (FSI), denotada por $S_{\text{pop}}(t)$ é expressa como:

$$\begin{aligned} S_{pop}(t) &= p_0 + (1-p_0)\int_t^{\infty} f(u)du \\ &= p_0 + (1-p_0)S(t), \quad 0 \leq p_0 \leq 1. \end{aligned} \tag{2.47}$$

Se $p_0 = 0$, então $S_{\text{pop}}(t) = S(t)$, ou seja, a função de sobrevivência. Além disso, a função de sobrevivência imprópria tem as seguintes propriedades:

1. $S_{\text{pop}}(0) = 1$;

2. $S_{\text{pop}}(t)$ é decrescente;

3. $\lim_{t\to\infty} S_{\text{pop}}(t) = p_0$, sendo essa a fração de cura.

A última propriedade retrata o fato da função de sobrevivência populacional (2.47) 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. É possível notar um exemplo dessa estabilização na Figura 2.6.

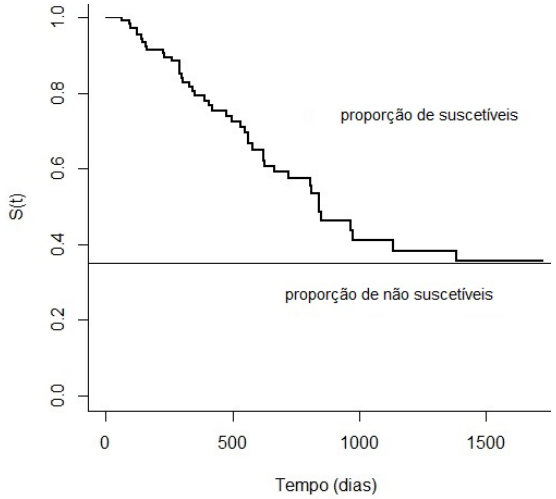

Figura 2.6: Curva de sobrevivência do modelo de fração de cura.

O modelo de mistura padrão permite lidar com populações heterogêneas, ou seja, compostas por diferentes subgrupos, cada um seguindo uma distribuição de sobrevivência distinta (Klein e Moeschberger, 2003). Nesse modelo, a população é dividida em subgrupos, e para cada subgrupo, é ajustada uma distribuição de sobrevivência específica. Essas distribuições são então combinadas para formar a distribuição geral da população. Esse modelo é utilizado em diversas áreas de pesquisa, como epidemiologia, ciências médicas e engenharia, especialmente quando se lida com dados que apresentam uma grande variabilidade entre os indivíduos ou quando existe a suspeita de diferentes comportamentos de sobrevivência na população.

O modelo de mistura padrão considera uma variável não observável de *Bernoulli* M_i. Essa variável indica se os indivíduos na amostra estão curados ou não, ou seja,

$$M_i = \begin{cases} 0, & \text{se o indivíduo } i \text{ está curado,} \\ 1, & \text{se o indivíduo } i \text{ está em risco,} \end{cases} \qquad i = 1, 2, ..., n. \tag{2.48}$$

Assim, $P(M_i = 0) = p_0$ e $P(M_i = 1) = 1 - p_0$. Desta forma, considerando T uma variável aleatória não negativa e contínua, representando o tempo de vida, sabe-se que

$$P(T > t|M_i = 1) = S(t) \tag{2.49}$$

e

$$P(T > t|M_i = 0) = 1. \tag{2.50}$$

A probabilidade do tempo de vida ser maior que um determinado tempo t, independente do grupo a que ele pertença, é dada por

$$\begin{aligned} S_{pop}(t) &= P(T > t) \\ &= P(T > t|M_i = 0)P(M_i = 0) + P(T > t|M_i = 1)P(M_i = 1) \\ &= p_0 + (1 - p_0)S(t), \quad t \geq 0. \end{aligned} \tag{2.51}$$

Assim, a função de sobrevivência populacional com proporção de curados é dada por 2.47, ou seja,

$$S_{pop}(t) = p_0 + (1 - p_0)S(t), \quad t \geq 0, \tag{2.52}$$

em que $S(.)$ representa a função de sobrevivência própria associada aos indivíduos em risco. A função de sobrevivência populacional $S_{pop}(t)$ (2.47) possui algumas propriedades, listadas anteriormente, e a última propriedade destaca que ela é imprópria. Isso ocorre porque a curva de sobrevivência estabiliza em p_0, que representa a probabilidade de cura da população

A função de densidade imprópria pode ser obtida a partir da derivação da função de sobrevivência imprópria,

$$f_{pop}(t) = -\frac{\partial[S_{pop}(t)]}{\partial t} = (1 - p_0)f(t), \tag{2.53}$$

e a função de risco populacional é definida como:

$$h_{pop}(t) = \frac{f_{pop}}{S_{pop}} = \frac{(1 - p_0)f(t)}{p_0 + (1 - p_0)S(t)}. \tag{2.54}$$

A partir das equações (2.51 e 2.54) é possível encontrar a função de risco própria,

$$h(t) = \frac{S_{pop}(t)h_{pop}(t)}{(1-p_0)S(t)} = \left[\frac{S_{pop}(t)}{S_{pop}(t)-p_0}\right] h_{pop}(t), \tag{2.55}$$

que relaciona a taxa de falha total na população com a taxa de falha própria dos indivíduos não curados, levando em conta a proporção de curados na população.

Função de Verossimilhança para o Modelo de Mistura Padrão

Considerando o conjunto de dados observados $D = (t, \delta)$, em que $t = (t_1, t_2, \ldots, t_n)'$ e $\delta = (\delta_1, \delta_2, \ldots, \delta_n)'$ representam os tempos de falha e os indicadores de falha/censura, respectivamente, a função de verossimilhança para o modelo de mistura padrão, com θ como o vetor de parâmetros a ser estimado, é dada por:

$$\begin{aligned} L(\theta; D) &\propto \prod_{i=1}^{n} [f_{pop}(t_i;\theta)]^{\delta_i} [S_{pop}(t_i;\theta)]^{1-\delta_i} \\ &\propto \prod_{i=1}^{n} [(1-p_0)f(t_i;\theta)]^{\delta_i} [p_0 + (1-p_0)S(t_i;\theta)]^{1-\delta_i}. \end{aligned} \tag{2.56}$$

Modelo de Longa Duração Exponencial

O modelo de mistura padrão exponencial combina duas subpopulações, uma que segue uma distribuição exponencial e outra que representa os indivíduos curados ou imunes ao evento de interesse. Esse modelo é derivado considerando uma variável *Bernoulli* não observável para os indivíduos curados e não curados na amostra. A distribuição exponencial é caracterizada por uma variável aleatória contínua não negativa, representada por T, que possui função densidade de probabilidade, função de sobrevivência e de risco, conforme descrito na seção 2.2.1.

A função de sobrevivência populacional e a função de densidade populacional, considerando a distribuição exponencial, são $S_{pop}(t)$ e $f_{pop}(t)$, respectivamente:

$$S_{pop}(t) = p_0 + (1-p_0)S(t) = p_0 + (1-p_0)e^{-\lambda t}, \tag{2.57}$$

$$f_{pop}(t) = (1-p_0)f(t) = (1-p_0)\lambda e^{-\lambda t}. \tag{2.58}$$

Portanto, a função de verossimilhança para o modelo de mistura padrão exponencial

é definida como:

$$\begin{aligned} L(\lambda|t,\delta) &= \prod_{i=1}^{n}[f_{pop}(t_i|\lambda)]^{\delta_i}[S_{pop}(t_i|\lambda)]^{1-\delta_i} \\ &= \prod_{i=1}^{n}[(1-p_0)\lambda e^{-\lambda t_i}]^{\delta_i}[p_0+(1-p_0)e^{-\lambda t_i}]^{1-\delta_i}, \end{aligned} \tag{2.59}$$

e consequentemente, a função de log-verossimilhança será definida por,

$$l(\lambda|t,\delta) = \sum_{i=1}^{n}\left[\delta_i \log((1-p_0)\lambda e^{-\lambda t_i}) + (1-\delta_i)\log(p_0+(1-p_0)e^{-\lambda t_i})\right]. \tag{2.60}$$

2.5.2 Aplicação do Modelo de Mistura Padrão

Com o intuito de aprofundar a compreensão sobre os modelos de mistura padrão, será realizada uma aplicação a um conjunto de dados, proveniente do pacote `Survival` do *Software R*, que contém informações sobre a sobrevivência de pacientes com câncer de cólon e possui as variáveis descritas na Tabela 2.1.

Tabela 2.1: Descrição das variáveis do conjunto de dados `colon`.

Variável	**Descrição**
Study	Estudo: 1 para todos os pacientes
Rx	Tratamento: Obs, Lev, Lev+5-FU
Sex	Sexo: 1 = masculino
Age	Idade: em anos
Obstruct	Obstrução do cólon pelo tumor
Perfor	Perfuração do cólon
Adhere	Adesão a órgãos próximos
Nodes	Número de linfonodos com câncer detectável
Time	Dias até o evento ou censura
Status	Status de censura
Differ	Diferenciação do tumor (1=bem, 2=moderado, 3=ruim)
Surg	Tempo da cirurgia até o registro (0=curto, 1=longo)
Node4	Mais de 4 linfonodos positivos
Etype	Tipo de evento: 1=recorrência, 2=morte
Extent	Extensão da disseminação local (1=submucosa, 2=músculo, 3=serosa, 4=estruturas contíguas)

Este conjunto de dados foi originalmente descrito em Laurie *et al.* (1989). Estes são dados de um dos primeiros ensaios bem-sucedidos de quimioterapia adjuvante para câncer de cólon. Dentre as 1.858 observações no conjunto de dados, 938 foram censuradas, representando aproximadamente 50,5% do total. O período de acompanhamento do estudo

foi pouco mais de 9 anos.

A Figura 2.7 apresenta a curva de sobrevivência, estimada por meio do estimador proposto por Kaplan e Meier (1958), observa-se que a curva de sobrevivência começa a estabilizar em torno de $S(t) = 0.45$, indicando que cerca de 45% dos pacientes apresentam resistência ao evento de interesse. Com isso, será realizado o ajuste do modelo de mistura padrão exponencial.

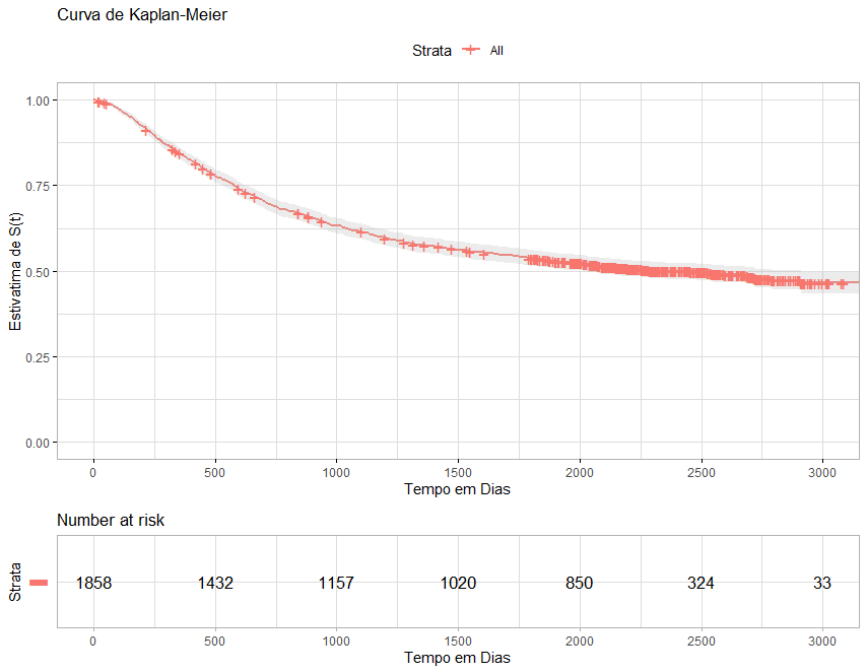

Figura 2.7: Curva de sobrevivência estimada através de *Kaplan-Meier*.

Os parâmetros do modelo são estimados por meio de uma maximização direta da função apresentada em (2.60), utilizando um método computacional apropriado. Dessa forma, a Tabela 2.2 apresenta os parâmetros estimados juntamente com os intervalos de confiança (IC 95%) para o modelo de mistura padrão exponencial.

Tabela 2.2: Estimativa de máxima verossimilhança (EMV) e intervalo de confiança - IC(95%) para o modelo exponencial.

Parâmetros	EMV	IC (95%)	
		LI	LS
λ	3,461	3,108	3,822
p_0	0,455	0,422	0,479

Para avaliar o ajuste do modelo aos dados, foi plotado um gráfico *Kaplan-Meier* (Figura 2.8). Os resultados indicam que a curva estimada do modelo exponencial ajustado é uma boa aproximação aos dados, isso se evidencia pela proximidade da curva da mo-

delagem ajustada com a curva do estimador de *Kaplan-Meier*, considerando também a normalização do tempo, que é importante quando os valores de tempo são muito grandes.

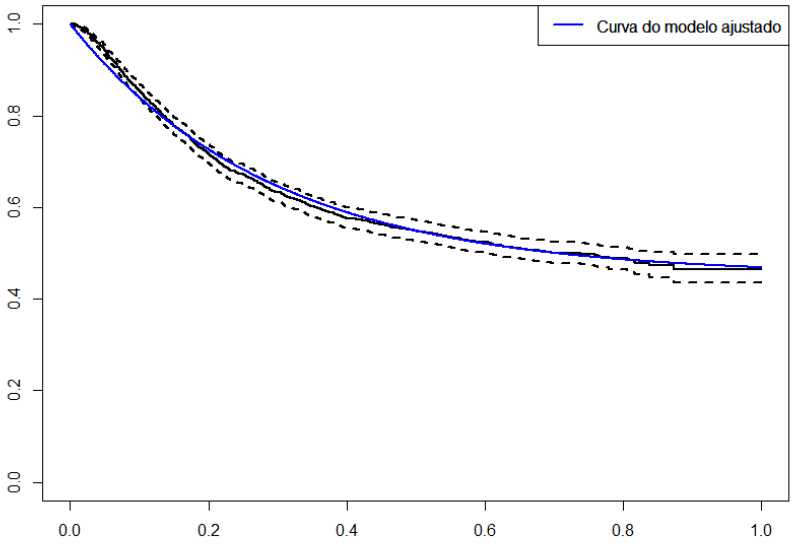

Figura 2.8: Comparação da curva de *Kaplan-Meier* com a curva do modelo exponencial ajustado.

A normalização ajusta a escala dos tempos para um intervalo específico, comumente entre 0 e 1, visando melhorar a modelagem, isso evita distorções nos resultados do modelo. No caso mencionado, os tempos foram divididos pelo maior valor de tempo observado, resultando em uma escala normalizada que varia entre 0 e 1.

2.6 Modelos 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 (Vaupel *et al.*, 1979).

O modelo de fragilidade engloba duas fontes de variação: 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.

A ideia básica de fragilidade (ou heterogeneidade não observada) é um fator de proporcionalidade aleatório não observado que modifica a função de risco de um indivíduo ou indivíduos relacionados. Quando os tempos de vida são univariados e independentes, a fragilidade pode ser usada para ajustar algum fator de risco não observado em um modelo de risco. Por outro lado, para tempos de vida multivariados e dependentes, a introdução de um efeito aleatório comum (fragilidade) é uma forma de modelar a dependência dos tempos dos eventos (Wienke, 2010). A fragilidade é introduzida na função de risco de forma multiplicativa ou aditiva.

2.6.1 Modelo de Fragilidade Multiplicativo para Dados Univariados

O modelo de fragilidade multiplicativo é uma extensão do modelo de *Cox*. No modelo de *Cox* (2.63), o risco de um evento ocorrer é modelado como uma função da combinação linear das variáveis explicativas, multiplicada por um fator comum a todos os indivíduos. Esse fator comum é a fragilidade, que captura as características não observadas que podem influenciar o risco de eventos. No entanto, o modelo de fragilidade multiplicativo vai além, ele incorpora essa fragilidade não observável de forma multiplicativa na função de risco, permitindo que ela tenha um efeito proporcional sobre o risco individual. Isso significa que a presença da fragilidade afeta o risco de maneira proporcional para todos os indivíduos, e não apenas através de um termo aditivo. A função de risco no instante t para o i-ésimo indivíduo, sem a influência de covariáveis, pode ser expressa como:

$$h_i(t|v) = h_0(t)v_i \quad i = 1, ..., n. \tag{2.61}$$

Nesta equação, $h_0(t)$ representa a função de risco base, que é comum a todos os indivíduos, e v_i é uma variável aleatória não negativa, independente e identicamente distribuída (i.i.d.).

O termo v_i é a medida de fragilidade do indivíduo, que indica o quão vulnerável ele é em relação ao evento de interesse. Quanto maior o valor de v_i, mais frágeis são as observações associadas ao i-ésimo indivíduo, por isso essa variável é chamada de fragilidade. Em outras palavras, quanto maior o valor de v_i, maior é a predisposição do indivíduo para experimentar o evento em questão. Portanto, é esperado que o evento de interesse ocorra com maior probabilidade para os indivíduos que apresentam valores mais elevados

de fragilidade, o que reflete a ideia de que a fragilidade está relacionada a uma maior suscetibilidade ao evento.

A função de sobrevivência condicional associada a esse modelo representa a probabilidade de um indivíduo estar vivo no tempo t dado o efeito aleatório v_i. Essa função pode ser definida da seguinte forma:

$$S_i(t|v_i) = [S_0(t)]^{v_i}, \tag{2.62}$$

em que $S_0(t)$ é a função de sobrevivência base, para toda a população, e representa a probabilidade de um indivíduo não ter experimentado o evento de interesse até o tempo t e a variável v_i atua como um multiplicador que modifica essa probabilidade.

O modelo de fragilidade multiplicativo na presença de covariáveis é:

$$h(t|v_i, x_i) = v_i h_0(t) exp\left\{\mathbf{x_i'}\boldsymbol{\beta}\right\}, \tag{2.63}$$

em que:

- v_i: representa a fragilidade do i-ésimo indivíduo;
- h_0: representa a função de risco base;
- $\boldsymbol{\beta}$: vetor de coeficientes a serem estimados;
- $\mathbf{x_i'}$: as covariáveis associadas ao i-ésimo indivíduo.

O modelo de fragilidade assume a 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 quando $v_i < 1$ e se reduz ao modelo de risco proporcional de *Cox* (2.23) quando $v_i = 1$.

O fato de que a variável de fragilidade atua 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. Portanto, é esperado que o evento de interesse ocorra para os indivíduos mais "frágeis".

2.6.2 Transformada de Laplace

A transformada de Laplace é fundamental nos modelos de fragilidade, a aplicação desse método em modelos de fragilidade é uma abordagem para obter a função marginal,

permitindo a estimativa dos parâmetros. Isso se torna necessário, uma vez que o método de máxima verossimilhança não suporta diretamente o uso de funções condicionadas.

A transformada de Laplace de uma função $g(.)$, considerada para um argumento s real, é definida como:

$$\mathcal{L}(s) = \int_0^\infty g(s)e^{-sx}ds, \tag{2.64}$$

sendo $\mathcal{L}[g(.)]$ a transformada de Laplace da função $g(.)$. Com essa definição, é possível obter a função de sobrevivência não condicional a partir da função de sobrevivência condicionada à fragilidade (2.62) e da função densidade da variável fragilidade, conforme a seguinte relação:

$$S(t) = \int_0^\infty S(t|v)g(v)dv = \int_0^\infty [S_0(t)]^v g(v)dv, \tag{2.65}$$

em que $S(t)$ se refere à função de sobrevivência não condicional, $S_0(t)$ representa a função de sobrevivência transformada e $g(v)$ a função densidade de probabilidade da variável de fragilidade. Logo,

$$S(t) = \int_0^\infty e^{-H_0(t)v}g(v)dv = \mathcal{L}[H_0(t)], \tag{2.66}$$

sendo que $\mathcal{L}[H_0(t)]$ denota a transformação de Laplace da função $g(v)$ considerando a função de risco acumulada transformada $H_0(t)$.

Além disso, esse método permite a obtenção de outros resultados relacionados à função de sobrevivência não condicional, considerando as derivadas da transformação (Wienke, 2007).

- **Função de densidade:**

$$f(t) = -h_0(t)\mathcal{L}'(H_0(t)). \tag{2.67}$$

- **Função de risco:**

$$h(t) = -h_0(t)\frac{\mathcal{L}'(H_0(t))}{\mathcal{L}(H_0(t))}. \tag{2.68}$$

- **Esperança da variável não observada V:**

$$E(V) = -\mathcal{L}'(0). \tag{2.69}$$

- **Variância da variável não observada V:**

$$Var(V) = \mathcal{L}''(0) - (\mathcal{L}'(0))^2, \tag{2.70}$$

em que $\mathcal{L}'$ e $\mathcal{L}''$ denotam a primeira e a segunda derivada da transformação de Laplace. A função de densidade dos sobreviventes, ou seja, dos indivíduos que não falharam até o instante t, é representada como:

$$f(v|T \geq t) = \frac{f(v)S(t|v)}{S(t)}, \tag{2.71}$$

em que $f(v)$ é a função densidade da variável de fragilidade v.

2.6.3 Distribuição Fragilidade Gama

A escolha da distribuição para a variável de fragilidade é uma etapa crítica, pois desempenha um papel fundamental na precisão das estimativas, nas inferências estatísticas e nas interpretações dos resultados obtidos. Diferentes distribuições podem ser atribuídas à variável de fragilidade, tais como gama, normal, log-normal, gaussiana inversa, Weibull, entre outras. A distribuição gama se destaca para modelar fragilidade devido à sua manipulação algébrica facilitada. Essa abordagem foi explorada por Vaupel *et al.* (1979). Além disso, em Tomazella (2003), é feita uma análise comparativa das distribuições gama, log-normal e gaussiana inversa, conduzida no contexto de modelos de regressão em processos de Poisson. Segundo Elbers e Ridder (1982), ao trabalhar com fragilidade é necessário que a distribuição do efeito aleatório tenha média finita para o modelo ser identificável. Sendo assim, a função densidade da distribuição gama definida em 2.16, considerando a seguinte parametrização $\gamma = 1$, $k = \alpha$ e $\lambda = 1/\alpha$, temos que $V \sim gama(\alpha, 1/\alpha)$, é expressa por

$$g(v) = \frac{\alpha^\alpha}{\Gamma(\alpha)} v^{\alpha-1} \exp\{-\alpha v\}, \tag{2.72}$$

em que $\alpha > 0$, sendo $E(V) = 1$ e $Var(V) = 1/\alpha$.

Especificamente nos modelos de fragilidade, o parâmetro α da distribuição gama é utilizado para modelar a heterogeneidade não observada entre os indivíduos. Quanto menor o valor de α, maior será a heterogeneidade, ou seja, os membros da população são mais semelhantes entre si em relação à característica ou resultado de interesse. Consequentemente, quando α é grande implica em pouca variabilidade entre os indivíduos, pois a

$Var(V) = 1/\alpha$ quantifica a heterogeneidade não observável entre os indivíduos.

Função de Sobrevivência e Risco não condicional

Quando a distribuição gama é utilizada para modelar a fragilidade, a partir da transformada de Laplace (2.64), encontramos a função de sobrevivência não condicional à variável fragilidade, dada por

$$\begin{aligned} S(t) &= \int_0^\infty \frac{\alpha^\alpha}{\Gamma(\alpha)} e^{-H_0(t)v} v^{\alpha-1} e^{-\alpha v} dv \\ &= \frac{\alpha^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-(H_0(t)v)} v^{\alpha-1} e^{-\alpha v} dv \\ &= \frac{\alpha^\alpha}{\Gamma(\alpha)} \int_0^\infty v^{\alpha-1} e^{-(H_0(t)+\alpha)v} dv \\ &= \frac{\alpha^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(H_0(t)+\alpha)^\alpha} = \left(\frac{\alpha}{H_0(t)+\alpha}\right)^\alpha \\ &= \left(\frac{1}{1+\frac{H_0(t)}{\alpha}}\right)^\alpha = \left[1+\frac{H_0(t)}{\alpha}\right]^{-\alpha}, \end{aligned} \tag{2.73}$$

com função densidade não condicional:

$$f(t) = h_0(t)\left[1+\frac{H_0(t)}{\alpha}\right]^{-\alpha-1}. \tag{2.74}$$

A função de risco não condicional é expressa como:

$$h(t) = \frac{\alpha h_0(t)}{\alpha + H_0(t)}. \tag{2.75}$$

É perceptível que a variável não observada de fragilidade V não se encontra nas equações 2.73 e 2.75. Consequentemente, as funções de sobrevivência e risco passam a depender apenas do parâmetro α a ser estimado. Considerando a função de risco de base exponencial dada em (2.13), com $H_0(t) = \lambda t$, a função de sobrevivência (2.73) e função de risco não condicional (2.75) são, respectivamente,

$$S(t) = \left[1+\frac{\lambda t}{\alpha}\right]^{-\alpha} \tag{2.76}$$

e

$$h(t) = \frac{\alpha\lambda}{\alpha+\lambda t}. \tag{2.77}$$

Capítulo 3

Modelo de Mistura Padrão com Fragilidade Gama

O modelo de mistura padrão (MMP) apresenta vantagens significativas em relação aos modelos de sobrevivência convencionais, uma vez que incorpora a heterogeneidade existente entre duas subpopulações distintas: aqueles que são curados e aqueles que não são. Ao modelar dados de sobrevivência que possuem uma longa duração, a inclusão da fragilidade no modelo é crucial para compreender variáveis não observáveis.

Considere a função de sobrevivência populacional (2.47), em que $S(t) = \mathcal{L}[H_0(t)]$ é obtida utilizando a transformada de Laplace (2.64). Então a função de sobrevivência populacional do modelo de mistura padrão com fragilidade é dada por

$$S_{pop}(t) = p_0 + (1 - p_0)\mathcal{L}[H_0(t)], \tag{3.1}$$

quando $p_0 = 0$, $S_{pop}(t) = \mathcal{L}[H_0(t)]$.

Assumindo que o componente representativo da fragilidade, denotado por V, segue uma distribuição gama ($V \sim gama(\alpha, \alpha)$), é possível reformular a função de sobrevivência populacional (3.1). Considerando a relação $\mathcal{L}[H_0(t)] = S(t)$ descrita em 2.73:

$$S_{pop}(t) = p_0 + (1 - p_0)\left(1 + \frac{H_0(t)}{\alpha}\right)^{-\alpha}, \tag{3.2}$$

em que,

- se $p_0 = 0$ e $\alpha \to \infty$, tem-se o modelo de sobrevivência usual;
- se $\alpha \to \infty$, tem-se o modelo de mistura padrão;

- se $p_0 = 0$, tem-se o modelo de fragilidade gama.

A função de densidade da população do modelo de mistura padrão com fragilidade gama é dada por:

$$f_{pop}(t) = (1 - p_0)\left(1 + \frac{H_0(t)}{\alpha}\right)^{-\alpha-1} h_0(t). \tag{3.3}$$

Utilizando a distribuição exponencial para modelar a taxa de risco da população, sendo a função de risco base denotada em 2.15. A função de sobrevivência populacional pode ser reescrita como

$$S_{pop}(t) = p_0 + (1 - p_0)\left(1 + \frac{\lambda t}{\alpha}\right)^{-\alpha}. \tag{3.4}$$

Dessa forma, a expressão para a função de densidade da população é dada por:

$$f_{pop}(t) = (1 - p_0)\left(1 + \frac{\lambda t}{\alpha}\right)^{-\alpha-1} \lambda. \tag{3.5}$$

Considerando a função de sobrevivência populacional, definida em 4.1, o modelo de mistura padrão com fragilidade gama sem covariáveis pode ser ajustado aos dados.

3.1 O modelo na presença de covariáveis observadas

O modelo de mistura padrão com fragilidade na presença de covariáveis observadas é utilizado para investigar como diferentes fatores ou características influenciam a taxa de falha ou sobrevivência de indivíduos em um estudo.

Assumindo uma distribuição gama para a fragilidade e considerando uma abordagem semiparamétrica, em que não foi atribuída uma distribuição de probabilidade para a função de risco base, Peng e Zhang (2008) incluíram cováriaveis nos componentes do modelo e estimaram os parâmetros de forma clássica, a partir do algoritmo EM (*Expectation-Maximization*) e o método de múltipla imputação.

A extensão do modelo de Berkson e Gage (1952) inclui o efeito de cováriavies e tem a função de sobrevivência populacional dada por:

$$S_{pop}(t|\mathbf{x}, \mathbf{z}) = p_0(\mathbf{z}) + (1 - p_0(\mathbf{z}))S(t|\mathbf{x}), \tag{3.6}$$

em que $\mathbf{z} = (z_1, z_2, ..., z_q)'$ representa as covariáveis que afetam a probabilidade de cura e

$\mathbf{x} = (x_1, x_2, ..., x_p)'$ as covariáveis que afetam a função de sobrevivência dos não curados, ambos conjuntos podem apresentar covariáveis em comum.

Quando uma distribuição de probabilidade é atribuída a t, o modelo torna-se paramétrico e para modelar o efeito das covariáveis será utilizada a função de ligação logito (Peng e Zhang (2008)), em que $\mathbf{b} = (b_0, b_1, ..., b_q)'$ é o vetor de parâmetros que serão estimados para as covariáveis associadas à fração de cura:

$$p_0(\mathbf{z}) = \frac{\exp(\mathbf{b}^T\mathbf{z})}{1 + \exp(\mathbf{b}^T\mathbf{z})}. \tag{3.7}$$

Ao substituir as estimativas dos parâmetros, encontra-se a estimação da probabilidade de cura. Dessa maneira, a distribuição dos indivíduos em risco, considerando a fragilidade com distribuição gama ($V \sim gama(\alpha, \alpha)$) (2.72), é definida por:

$$S(t|\mathbf{x}) = \left(1 + \frac{H_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta})}{\alpha}\right)^{-\alpha}. \tag{3.8}$$

Consequentemente, a função de densidade e de risco com covariáveis, são, respectivamente:

$$f(t|\mathbf{x}) = \left(1 + \frac{H_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta})}{\alpha}\right)^{-\alpha-1} h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}) \tag{3.9}$$

e

$$h(t|\mathbf{x}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1 + \left(\frac{H_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta})}{\alpha}\right)} h_0(t), \tag{3.10}$$

em que, o efeito das covariáveis $\mathbf{x}$ não satisfaz a suposição de riscos proporcionais (Taconeli, 2013). A substituição de (3.8) em (3.6) resulta na função de sobrevivência populacional com fração de cura e fragilidade gama na presença de covariáveis, expressa por Kuk e Chen (1992), como:

$$S_{pop}(t|\mathbf{x}, \mathbf{z}) = p_0(\mathbf{z}) + (1 - p_0(\mathbf{z}))\left(1 + \frac{H_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta})}{\alpha}\right)^{-\alpha}, \tag{3.11}$$

em que, $\mathbf{x}$ e $\mathbf{z}$ são vetores de covariáveis associadas à função de sobrevivência dos não curados e à fração de cura, respectivamente. A função densidade da população associada a 3.11 é dada por

$$f_{pop}(t|\mathbf{x}, \mathbf{z}) = (1 - p_0(\mathbf{z}))\left(1 + \frac{H_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta})}{\alpha}\right)^{-\alpha-1} h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}). \tag{3.12}$$

3.2 Inferência do MMP com Fragilidade Gama

Para realizar a estimação, será utilizado o método de máxima verossimilhança e assim é necessária a obtenção da função de verossimilhança, para isso são considerados os dados observados, representados por $\mathcal{D} = (t, \delta)$, sendo t o tempo de ocorrência, δ a variável indicadora, que denota se o evento foi censurado ($\delta_i = 0$) ou não ($\delta_i = 1$), e o conjunto de parâmetros $\xi = (\alpha, b, \boldsymbol{\beta})$. Dessa maneira, a função de verossimilhança para o modelo de mistura padrão, é definida como:

$$L(\xi|\mathcal{D}) = \prod_{i=1}^{n} \left[S_{pop}(t_i;\xi)\right]^{1-\delta_i} \left[f_{pop}(t_i;\xi)\right]^{\delta_i} \tag{3.13}$$

.

Ao substituir as equações (3.11) e (3.12) na função de verossimilhança para o modelo de mistura padrão com covariáveis e fragilidade com distribuição gama, temos:

$$\begin{aligned} L(\xi|\mathcal{D}) = \prod_{i=1}^{n} &\left[p_0(\mathbf{z}_i) + (1-p_0(\mathbf{z}_i)) \left(1 + \frac{H_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{\alpha}\right)^{-\alpha}\right]^{1-\delta_i} \times \\ &\left[(1-p_0(\mathbf{z}_i)) \left(1 + \frac{H_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{\alpha}\right)^{-\alpha-1} h_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta})\right]^{\delta_i}. \end{aligned} \tag{3.14}$$

Considerando que as funções de risco base e risco acumulado seguem uma distribuição Weibull com parâmetros (λ, ν), conforme descrito na seção 2.2.4, a função de verossimilhança 3.14 pode ser reformulada como:

$$\begin{aligned} L(\xi|\mathcal{D}) = \prod_{i=1}^{n} &\left[p_0(\mathbf{z}_i) + (1-p_0(\mathbf{z}_i)) \left(1 + \frac{(t\lambda)^{\nu}\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{\alpha}\right)^{-\alpha}\right]^{1-\delta_i} \times \\ &\left[(1-p_0(\mathbf{z}_i)) \left(1 + \frac{(t\lambda)^{\nu}\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{\alpha}\right)^{-\alpha-1} \nu\lambda(t\lambda)^{\nu-1}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\right]^{\delta_i}, \end{aligned} \tag{3.15}$$

em que:

- α: parâmetro da distribuição do termo de fragilidade, gama;
- λ e ν: parâmetros da função de risco base com distribuição Weibull;
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$: parâmetros relativos às covariáveis.

Os parâmetros do modelo precisam ser estimados para que seja possível encontrar os valores que melhor se ajustam aos dados observados, essa estimativa é obtida ao maximizar

a função de verossimilhança (3.14), que mede a probabilidade dos dados sob o modelo. No entanto, devido à complexidade dessa função e às possíveis não linearidades, a otimização é executada utilizando a função `optim` do *software R*. Esse processo busca encontrar os valores de parâmetros que tornam os dados observados mais prováveis de acordo com o modelo escolhido.

3.3 Aplicação do MMP com Fragilidade Gama

O conjunto de dados deriva de um estudo envolvendo indivíduos diagnosticados com melanoma, conduzido com o propósito de avaliar a eficácia da administração de doses elevadas de interferon alfa-2b como medida preventiva contra a recorrência do câncer de pele. Os participantes foram recrutados para o estudo entre os anos de 1991 e 1995, sendo acompanhados até o ano de 1998. Mais detalhes do conjunto de dados pode ser obtido a partir da consulta do trabalho de Kirkwood e Austad (2000).

Para essa aplicação, foi feita a suposição que os dados seguem uma função de risco de base Weibull (2.21). No *Software R*, as funções foram definidas com base nos parâmetros (α, ν, λ e p_0) e a função `optim` foi utilizada para otimizar a função verossimilhança (3.15).

A Tabela 3.1 apresenta os resultados obtidos da Estimativa de Máxima Verosimilhança (EMV) dos parâmetros do modelo com função de risco base Weibull (3.3), juntamente com os intervalos de confiança - IC(95%), sendo "LI" o limite inferior e "LS" o limite superior. Além disso, são fornecidos os desvios padrões, que indicam a variabilidade das estimativas, os resultados do teste de *Wald* e os p-valores para avaliar a significância estatística dos parâmetros.

Os resultados dos testes apresentam valores extremamente baixos nos p-valores e altos no teste de *Wald*, isso indica a significância estatística para esses parâmetros. No entanto, α possui um p-valor mais alto, indicando que pode não ser estatisticamente significativo, ou seja, não é possível afirmar que α tem um papel significativo.

O valor estimado para p_0 é 0,449, indicando que a a proporção de cura é de 44,9% para esse conjunto de dados. A estimativa de $\nu > 1$, indica que a taxa de falha aumenta com o tempo, ou seja, a forma da função de risco tem um impacto estatisticamente significativo na recorrência do melanoma ao longo do tempo.

A Figura 3.1 mostra a comparação da curva de *Kaplan-Meier* com a curva do modelo ajustado. A concordância é notável pelo fato de que a curva do modelo ajustado prati-

Tabela 3.1: Resultados da análise dos dados de Kirkwood e Austad (2000) para o modelo (3.5).

Parâmetro	EMV	Desvio Padrão	IC (95%)		P-Valor	*Wald*
			LI	LS		
α	0,721	0,493	0,378	1,064	0,14	1,464
λ	1,764	0,182	1,638	1,891	0,000	9,708
ν	2,230	0,301	2,021	2,439	1,1e-13	7,422
p_0	0,449	0,071	0,400	0,499	2,4e-10	6,330

camente se sobrepõe à curva do estimador proposto por Kaplan e Meier (1958). Dessa maneira, os resultados reforçam a conclusão de que o modelo ajustado, que incorpora a função de risco base Weibull e a fragilidade gama, se ajustou bem ao conjunto de dados.

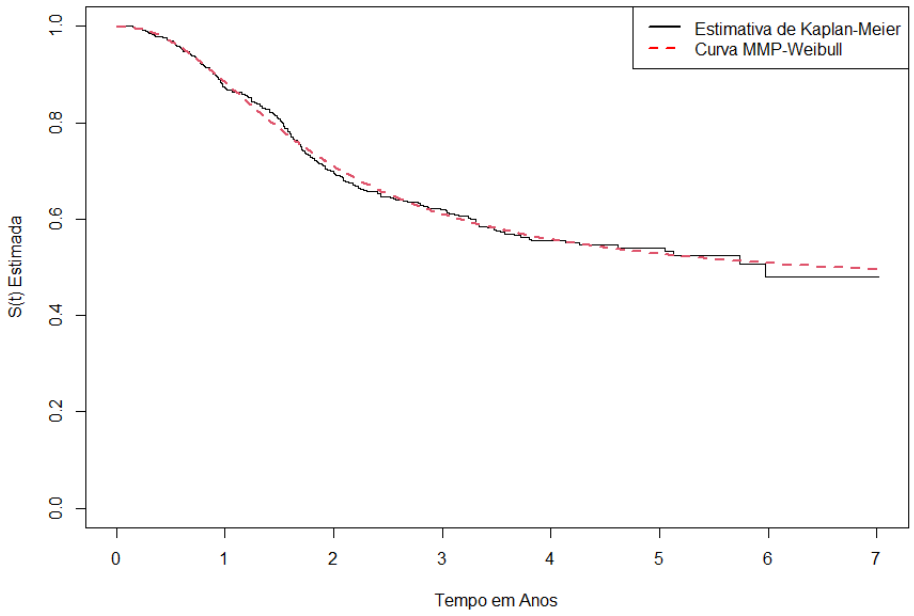

Figura 3.1: Comparação da curva de *Kaplan-Meier* com a curva do modelo de mistura padrão com fragilidade gama e função de risco base Weibull.

Capítulo 4

Aplicação aos Dados de Melanoma

Os dados aplicados neste trabalho foram coletados a partir de um estudo realizado com pacientes que receberam o diagnóstico de melanoma entre 2000 e 2014, com acompanhamento realizado até 2018, conduzido na Fundação Oncocentro de São Paulo (FOSP), que coordena o Registro Hospitalar de Câncer do Estado de São Paulo.

A FOSP é uma instituição pública ligada à Secretaria Estadual de Saúde, que auxilia na preparação e implementação de políticas de saúde no campo da Oncologia, e serve como um instrumento para que hospitais de oncologia possam elaborar seus próprios protocolos e melhorar suas práticas de cuidados. O melanoma é uma forma de câncer de pele que tem origem nas células produtoras de melanina, o pigmento responsável pela coloração da pele.

No escopo desse estudo, o foco foi investigar a ocorrência de óbitos relacionados ao melanoma, sendo esse o evento de interesse. Esse conjunto de dados foi abordado por outros autores, em que foram utilizadas outras metodologias, como modelagem de fragilidade de longo prazo usando um modelo de riscos não proporcionais por Calsavara *et al.* (2020) e modelos de fragilidade discreta de séries de potências com modificação zero por Molina *et al.* (2021).

O conjunto de dados inclui 7.823 registros, e a Tabela 4.1 apresenta as variáveis e suas respectivas descrições. A distribuição da variável sexo é quase igualmente dividida entre pacientes do sexo feminino e masculino. A idade dos participantes do estudo varia de jovens até idosos, com a idade máxima atingindo 99 anos. O período de estudo se estendeu por até 18,5 anos, e 73,4% dos participantes do estudo foram censurados, sendo esses os pacientes que não enfrentaram o evento de interesse por decorrência do melanoma e durante esse período foram caracterizados como observações com censura à direita.

Tabela 4.1: Descrição das variáveis.

Variável	**Descrição**
Idade	Idade do paciente
Sexo	Sexo (0 para masculino, 1 para feminino)
EC_cat	Estágio clínico (0 para estágios I e II, 1 para estágios III e IV)
Cirurgia	Cirurgia relacionada ao melanoma (0 para não, 1 para sim)
Radio	Radioterapia (0 para não, 1 para sim)
Quimio	Quimioterapia (0 para não, 1 para sim)
Tempo_anos	Tempo em anos até evento de interesse
Status	Censura (0 para censura, 1 para não censurados)

Conforme o site Onco Guia, o estágio é um método para descrever um câncer, incluindo sua localização, a presença de disseminação e o impacto nos órgãos do corpo. Compreender o estágio auxilia os médicos a determinar o tratamento adequado e o prognóstico do paciente. Calsavara *et al.* (2020) cita que de acordo com a última edição (3) do *American Joint Committee on Cancer* (AJCC), os estágios clínicos I e II correspondem ao melanoma limitado à pele, que está associado a um melhor prognóstico, o estágio clínico III corresponde à disseminação nodal do melanoma, em que nesse cenário a cirurgia é rotineiramente associada à radioterapia e/ou a alguma modalidade de tratamento sistêmico, e o estágio clínico IV corresponde à doença metastática, que possui o pior prognóstico. A partir das informações disponíveis sobre o estágio clínico, cerca de 68% dos pacientes estavam nos estágios clínicos I e II.

4.1 Análise Descritiva

Nesta seção, será realizada uma análise descritiva dos dados com o objetivo de oferecer um detalhamento das características das variáveis presentes no conjunto de dados. Entre os 7.823 pacientes, 26,6% enfrentaram o evento de interesse relacionado ao melanoma.

A Figura 4.1 apresenta a representação gráfica que destaca a variabilidade no tempo até o óbito (em anos) após o ingresso do paciente no estudo. Há uma concentração significativa de casos entre 0 e 5 anos, ao decorrer do tempo essa distribuição diminui, tendo os menores picos após 11 anos de estudo. No *boxplot*, a mediana é de aproximadamente 3 anos e é evidente a presença de valores discrepantes, também conhecidos como *outliers*. Assim como no histograma esses valores se concentram após os 12 anos desde a entrada no estudo, adicionalmente, o tempo até o óbito varia de 0 e 19 anos, em que 75% dos pacientes sobreviveram até cerca de 6 anos.

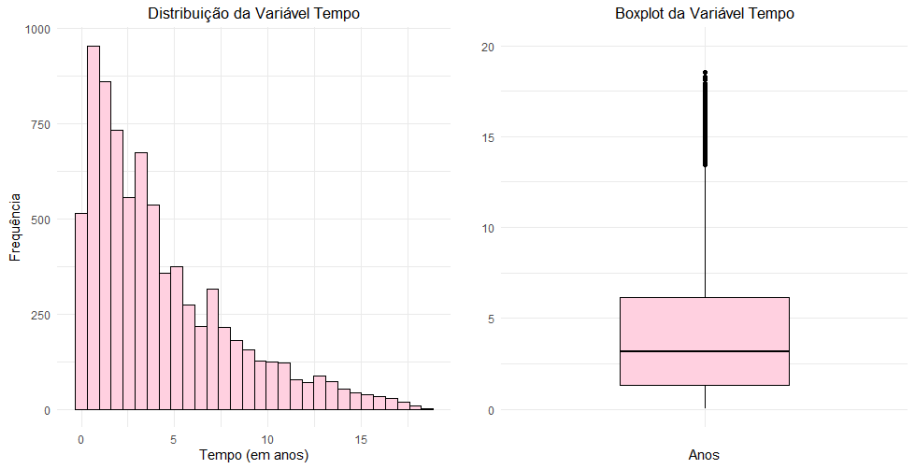

Figura 4.1: Histograma e *boxplot* da variável tempo até óbito.

A Figura 4.2 exibe informações sobre a idade do paciente (em anos). A análise revela que a mediana da idade dos pacientes é de aproximadamente 59 anos e que existem *outliers* nos dados, além disso 75% dos pacientes tem até 71 anos, sendo que a idade varia de 0 a 99 anos e se concentra entre 40 e 80 anos.

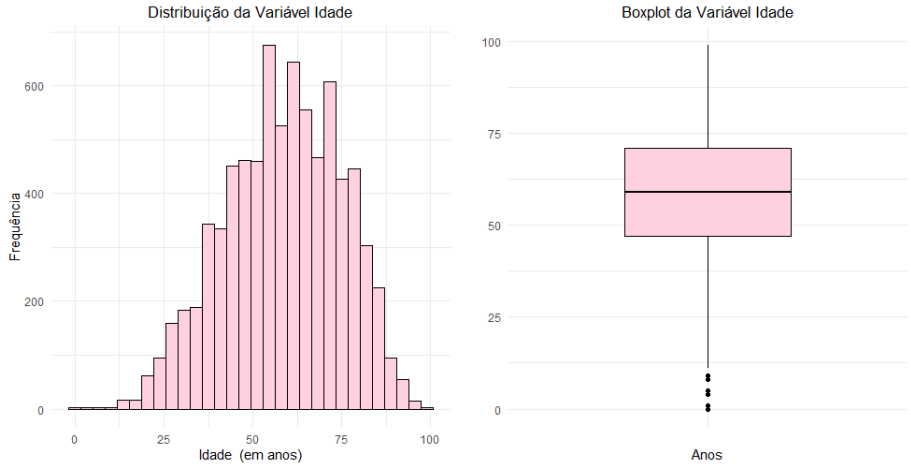

Figura 4.2: Histograma e *boxplot* da variável tempo idade do paciente.

A Tabela 4.2 auxilia e confirma as análises feitas nos gráficos das variáveis `tempo` e `idade`, algumas conclusões podem ser adicionadas. A variabilidade do tempo até o óbito sugere que alguns óbitos ocorreram logo no início do estudo, enquanto outros se estenderam por um período muito mais longo. A média, que é em torno de 4,23 anos, fica um pouco acima da mediana, indicando a possível influência de *outliers*. Essa diferença entre a média e a mediana pode ser explicada pelo fato de que alguns valores extremos (*outliers*) estão puxando a média para cima, enquanto a mediana, sendo menos sensível a extremos, mantém-se mais próxima da tendência central da maioria dos dados A dis-

tribuição da idade do paciente parece mais uniforme, já que a média e a mediana estão próximas, sugerindo uma distribuição relativamente equilibrada.

Tabela 4.2: Resumo das variáveis tempo até o óbito e idade.

Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Tempo	0,003	1,294	3,153	4,232	6,149	18,541
Idade	0,000	47,000	59,000	58,420	71,000	99,000

A Tabela 4.3 exibe a distribuição das variáveis clínicas presentes no estudo. A amostra demonstra uma distribuição quase equilibrada entre os pacientes do sexo masculino (49,51%) e feminino (50,49%). Os estágios do melanoma foram divididos em duas categorias (EC_cat), em que 68,17% dos pacientes estão nos estágios iniciais (I e II) e 31,83% nos estágios mais avançados (III e IV).

Tabela 4.3: Distribuição das variáveis EC_cat, sexo, cirurgia, radio e quimio.

Estágio		**Sexo**			**Cirurgia**	**Radio**	**Quimio**
I e II	5.331	**Masculino**	3.865	**Não**	916	7.157	6.579
em %	68,17	**em %**	49,51	**em %**	11,71	91,26	84,19
III e IV	2.492	**Feminino**	3.958	**Sim**	6.907	666	1.244
em %	31,83	**em %**	50,49	**em %**	88,29	8,74	15,81

Adicionalmente, quanto aos tratamentos, a maioria dos pacientes passaram por cirurgia (88,29%), enquanto a maioria não recebeu tratamento de radioterapia (91,26%) e quimioterapia (84,19%). As Figuras 4.3 e 4.4 permitem visualizar a distribuição das variáveis.

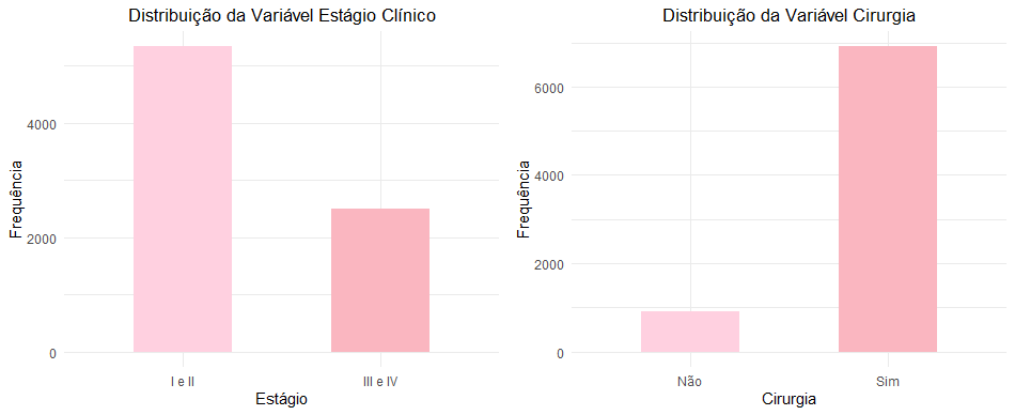

Figura 4.3: Histograma das variáveis discretas status, cirurgia, radioterapia e quimioterapia.

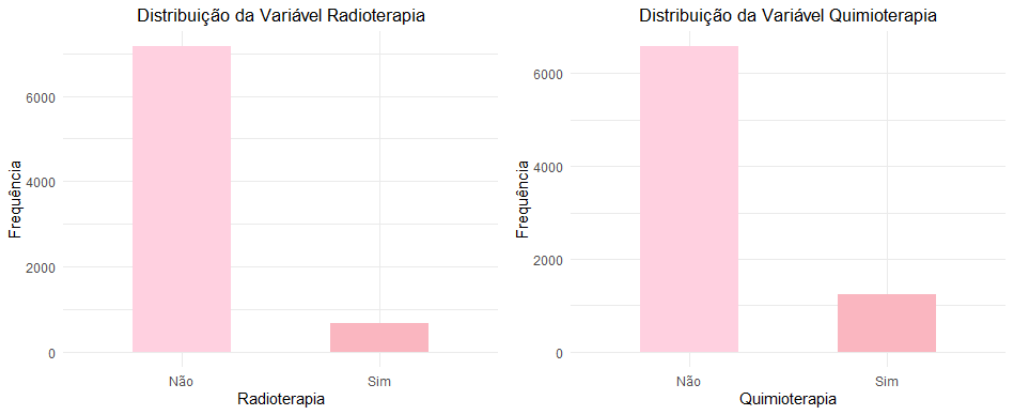

Figura 4.4: Histograma das variáveis discretas status, cirurgia, radioterapia e quimioterapia.

4.2 Estimador de Kaplan-Meier

Nesta seção, estão apresentadas as curvas de sobrevivência estimadas no *R* para diferentes grupos como sexo, realização de cirurgia, administração de quimioterapia e radioterapia. Essas análises permitem a visualização das diferenças nas taxas de sobrevivência entre os grupos, contribuindo para uma melhor compreensão do impacto de fatores específicos nas trajetórias de sobrevivência dos pacientes. O gráfico de *Kaplan Meier* do público geral na Figura 4.5 representa a função de sobrevivência estimada para a população total de pacientes, independentemente de quaisquer características específicas.

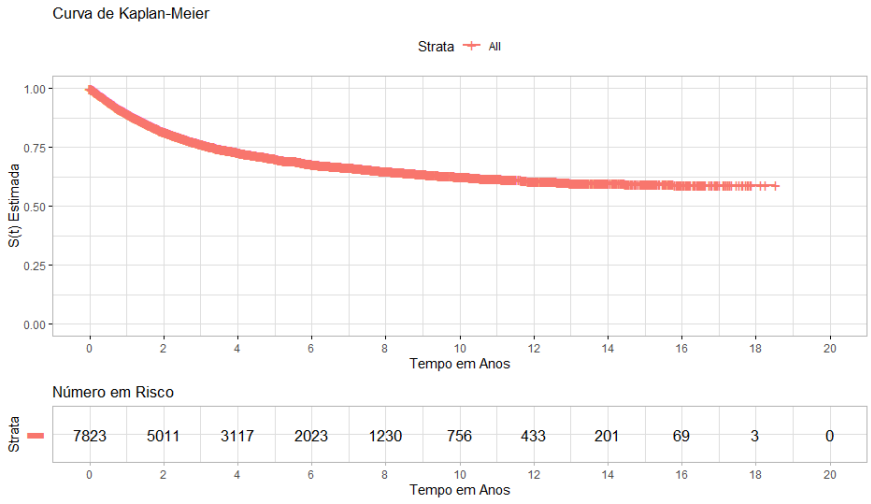

Figura 4.5: Curva de Sobrevivência estimada via Kaplan-Meier para o conjunto de dados de melanoma.

A curva mostra como a probabilidade de sobrevivência evolui ao longo do tempo e à medida que o tempo passa diminui gradualmente, o que indica uma redução na probabilidade de sobrevivência. Os traços na curva correspondem aos momentos em que

ocorrem eventos de interesse. No entanto, após um certo ponto, a curva tende a se estabilizar próximo a 0,6. Essa estabilização indica a "fração de cura", que representa os pacientes que superaram o risco de óbito referente ao melanoma durante o período de observação e são considerados curados.

Na Figura 4.6 foram identificadas discrepâncias nas curvas de sobrevivência para pacientes do sexo masculino e feminino, sugerindo possíveis diferenças nas taxas de sobrevivência entre os grupos. A curva de sobrevivência para o sexo feminino está acima da curva do sexo masculino, indicando uma taxa de sobrevivência melhor para as mulheres.

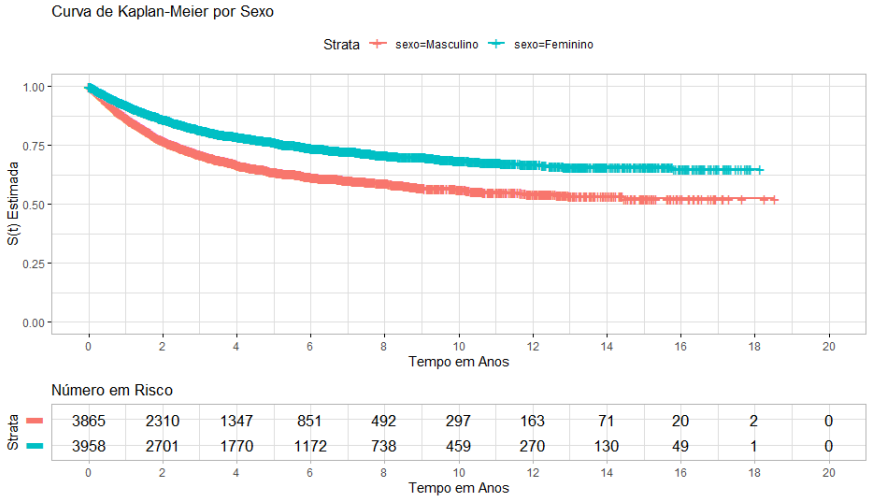

Figura 4.6: Função de sobrevivência estimada pelo *Kaplan-Meier* para o estrato Sexo.

Conforme a Sociedade Brasileira de Cirurgia Oncológica (2023), a excisão cirúrgica é uma das opções terapêuticas mais utilizadas para tratar o melanoma, seja maligno ou não. No caso das lesões benignas, embora seja um procedimento relativamente mais simples, essa é considerada uma providência para a cura. O procedimento consiste na remoção da lesão mais as margens laterais do tecido saudável, para que todas as células cancerosas sejam retiradas.

Na análise da Figura 4.7, pode-se notar que os pacientes submetidos à cirurgia apresentam uma taxa de sobrevivência maior em comparação com os que não realizaram procedimento cirúrgico. Isso ressalta a importância da cirurgia como parte do tratamento, sugerindo uma associação direta entre o procedimento cirúrgico e os melhores desfechos de sobrevida dos pacientes com melanoma.

Segundo o American Cancer Society (2023), o tipo de tratamento varia de acordo com o estágio do câncer, para os primeiros estágios a cirurgia é opção mais recomendada. Ao realizar a cirurgia (excisão ampla), se as bordas da amostra retirada não contiverem

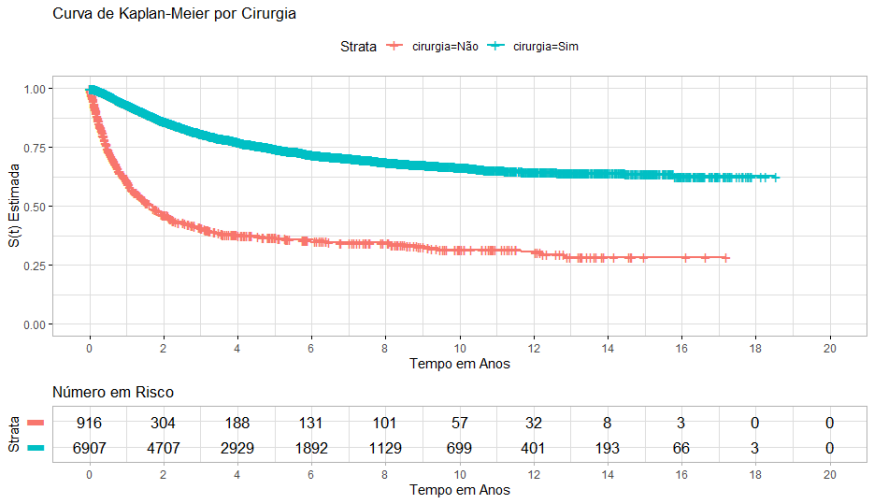

Figura 4.7: Função de sobrevivência estimada pelo *Kaplan-Meier* para o estrato Cirurgia.

células cancerígenas, basta fazer o acompanhamento adequado, porém em estágios mais avançados pode ser que não seja possível remover toda a área prejudicada, pois o melanoma se disseminou para os linfonodos e nesse caso é necessário realizar outros tipos de tratamento, como imunoterapia, radioterapia e quimioterapia.

A eficácia da radioterapia e quimioterapia no tratamento do melanoma pode variar dependendo do estágio do câncer, da extensão da doença e de outros fatores individuais. Em geral, o melanoma é conhecido por ser relativamente resistente à quimioterapia tradicional, pois tem um papel limitado no tratamento de melanomas avançados e muitas vezes é menos eficaz do que em alguns outros tipos de câncer.

Para os grupos que receberam o tratamento de quimioterapia (Quimio), a taxa de sobrevivência estimada foi menor em comparação aos grupos que não receberam esses tratamentos. Isso significa que os pacientes submetidos à quimioterapia tiveram uma probabilidade menor de sobrevivência ao longo do tempo em comparação com aqueles que não foram submetidos a esse tratamento, como é possível verificar na Figura 4.8. Isso se deve ao fato de que a quimioterapia é recomendada para estágios mais avançados do câncer, nos quais a probabilidade de sobrevivência já é menor, ainda podendo ser, apesar dos efeitos colaterais, uma opção menos agressiva, como amputação de membro em alguns casos.

A radioterapia também pode ter um papel limitado no tratamento do melanoma, especialmente se estiver localizado em áreas de difícil acesso para aplicação da radiação ou se o melanoma já se espalhou para outras partes do corpo (metástases). Assim como para a quimioterapia, na Figura 4.9 é possível notar que a taxa de sobrevivência estimada é

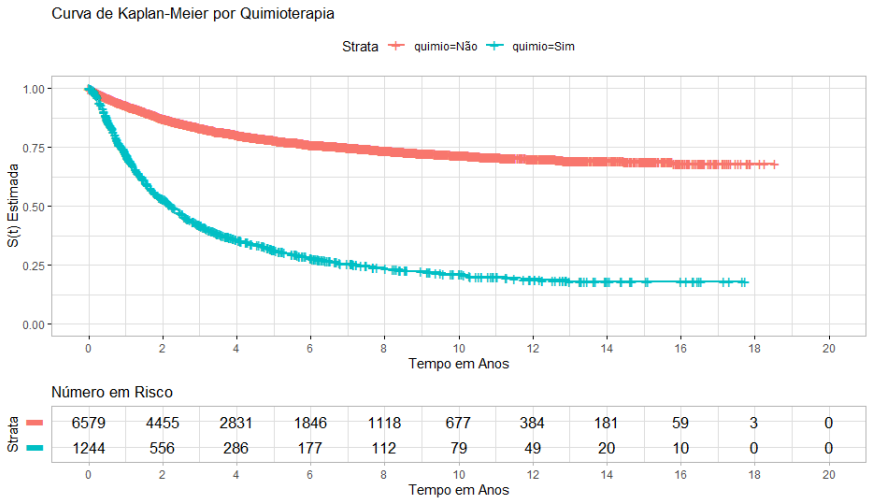

Figura 4.8: Função de sobrevivência estimada pelo *Kaplan-Meier* para o estrato Quimioterapia.

maior para os pacientes que não realizaram esse tratamento. A decisão sobre o tratamento ideal deve ser baseada em uma avaliação completa da situação médica, considerando tanto os aspectos de sobrevivência quanto de qualidade de vida.

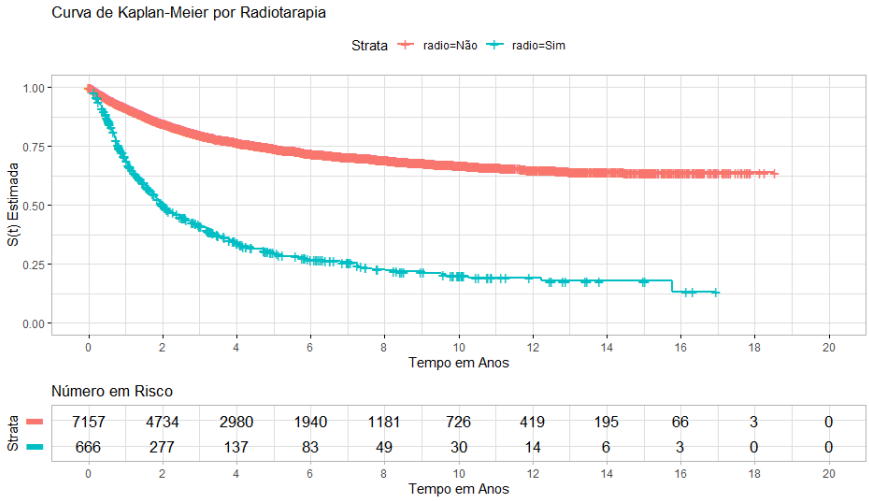

Figura 4.9: Função de sobrevivência estimada pelo *Kaplan-Meier* para o estrato Radioterapia.

Para a estimação da função de sobrevivência por estrato, foi utilizado o pacote "*survminer*" do *software R*. Este pacote não apenas facilita a análise estatística, mas também oferece vantagens na apresentação gráfica dos resultados. Sua abordagem gráfica aprimorada permite uma visualização mais clara e compreensível das diferenças nas curvas de sobrevivência entre os estratos. Essa escolha contribuiu para uma análise mais aprofundada nas taxas de sobrevivência entre os grupos de interesse.

4.3 Ajuste do Modelo de Mistura Padrão Exponencial

Nesta seção, será feita a aplicação do Modelo de Mistura Padrão com Fragilidade Gama visto no Capítulo 3. Conforme explorado anteriormente, esse modelo incorpora a flexibilidade da distribuição gama para capturar a heterogeneidade não observada entre as unidades experimentais. Serão realizadas abordagens distintas, considerando análises tanto sem covariáveis quanto com covariáveis, isso permitirá uma compreensão da influência de variáveis adicionais, enriquecendo a análise e proporcionando entendimentos sobre os fatores que podem modular a fragilidade.

4.3.1 Modelo sem Covariáveis

Considerando a função de sobrevivência populacional definida em 3.2, o modelo de mistura padrão com fragilidade gama sem covariáveis foi ajustado.

A Tabela 4.4 apresenta os valores obtidos. Os três parâmetros são estatisticamente significativos. O parâmetro λ tem uma estimativa de $0,292$ e conforme visto em 2.6.3, o parâmetro α é utilizado para modelar a heterogeneidade não observada entre os indivíduos, a estimativa de $\alpha = 2,737$ resulta na variância da distribuição de fragilidade ser igual a $Var(V) = 1/\alpha = 0,365$, a qual é significativa. O parâmetro p_0 tem uma estimativa de $0,565$, com um IC entre $0,525$ e $0,605$, refletindo a proporção de indivíduos considerados curados. Devido aos valores do teste *Wald* e do p-valor, há indícios da significância estatística de cada parâmetro.

Tabela 4.4: Estimativa de máxima verossimilhança (EMV), intervalo de confiança - IC(95%), desvio padrão (DP), teste de *Wald* e p-valor dos parâmetros para o modelo 3.2.

Parâmetro	EMV	IC (95%)		DP	*Wald*	P-Valor
		LI	LS			
λ	0,292	0,264	0,319	0,014	20,849	0,000
α	2,737	0,629	4,844	1,075	2,546	0,011
p_0	0,565	0,525	0,605	0,021	27,523	0,000

O ajuste do modelo é evidenciado na Figura 4.10, a sobreposição da curva estimada do modelo sobre a curva de Kaplan-Meier confirma que o modelo captura de maneira eficaz a heterogeneidade na população.

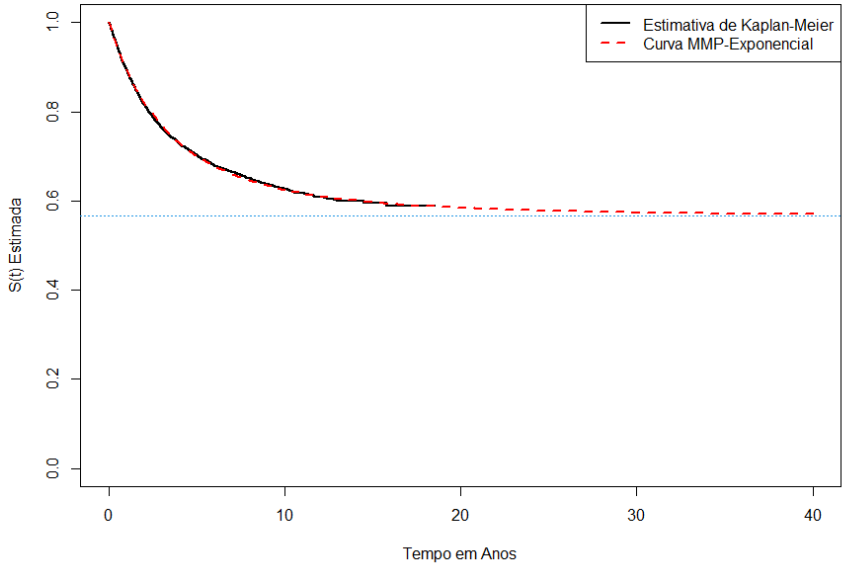

Figura 4.10: Curvas de sobrevivência para o modelo de mistura padrão com fragilidade gama e função de risco base exponencial sem a presença de covariáveis, a linha tracejada em azul representa a fração de cura.

4.3.2 Modelo com Covariáveis

Após realizar o ajuste inicial do modelo sem covariáveis, será apresentada uma análise para o modelo incluindo as covariáveis individuais na população de curados. O modelo ajustado se refere a equação (3.11), substituindo a função de risco base $H_0(t) = \lambda t$, a função de sobrevivência resultante é dada por:

$$S_{pop}(t) = p_0(\mathbf{x}) + (1 - p_0(\mathbf{x}))\left(1 + \frac{\lambda t}{\alpha}\right)^{-\alpha}. \tag{4.1}$$

Para estimar a proporção de cura $p_0(\mathbf{x})$ na equação 4.1, utiliza-se a função de ligação logito (3.7) dada por

$$p_0(\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}. \tag{4.2}$$

A covariável x_1 refere-se as covariáveis que serão incluídas no modelos de forma individual: sexo, estagio do câncer, realização de cirurgia referente ao melanoma, radioterapia ou quimioterapia.

- **Covariável Sexo**

  A Tabela 4.5 mostra as estimativas de máxima verossimilhança (EMV), o Intervalo

de confiança (IC), o Desvio Padrão (DP), o P-valor e a estatística de teste *Wald* para os parâmetros do modelo. Os parâmetros λ, α e β_1 são estatisticamente significativos com a inclusão da covariável sexo. A estimativa de $\alpha = 2,484$ resulta na variância da distribuição de fragilidade, em que $Var(V) = 1/\alpha = 0.403$, a qual é positiva e representa a heterogeneidade não observada de fatores externos. A estimativa positiva de β_1 indica que taxa de sobrevivência é maior para pacientes do sexo feminino (sexo=1).

Tabela 4.5: Resultados da análise dos dados considerando a covariável sexo no modelo (3.11).

Parâmetro	EMV	IC (95%)		DP	*Wald*	P-Valor
		LI	LS			
λ	0,287	0,259	0,316	0,014	19,996	0,000
α	2,484	0,684	4,284	0,918	2,705	0,007
β_0	-0,112	-0,326	0,101	0,109	-1,031	0,302
β_1	0,690	0,550	0,829	0,071	9,679	0,000

Na Figura 4.11 nota-se que o modelo proposto se ajustou bem aos dados, pois ao comparar a curva de sobrevivência estimada do modelo com a curva de *Kaplan-Meier*, é possível notar a proximidade entre elas. A proporção de cura estimada para o sexo feminino é $p_0 = 0,641$, enquanto para o sexo masculino foi de $p_0 = 0,472$, indicando que a população de pacientes do sexo feminino tem maior proporção de cura.

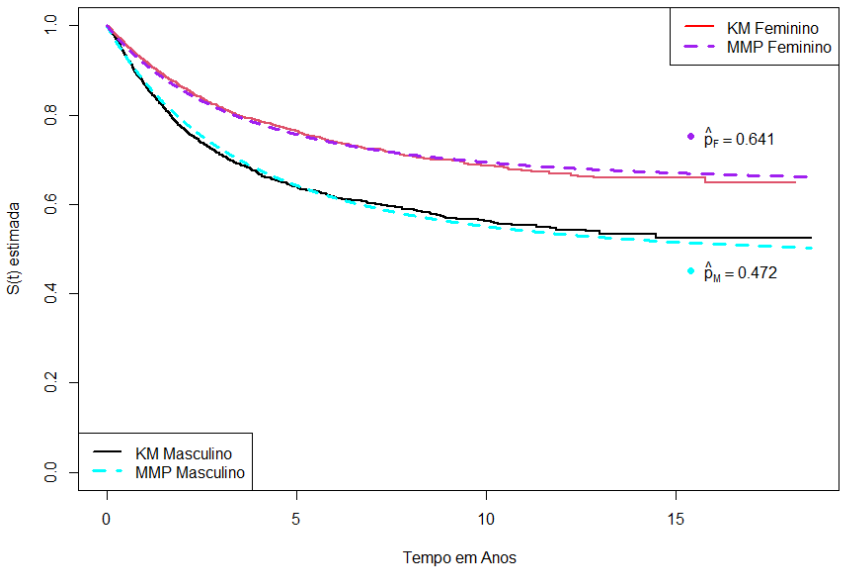

Figura 4.11: Curvas de sobrevivência para o modelo (3.11) com a covariável sexo.

- **Covariável Estágio**

  A análise da Tabela 4.6, destaca a significância estatística dos parâmetros estimados com a inclusão da covariável estágio. O parâmetro β_1, relacionado à covariável do estágio do melanoma, exibe p-valor muito baixo, sugerindo a influência estatisticamente significativa.

  O parâmetro α é o único que possui um p-valor um pouco mais alto (0,015), porém não alto suficiente para não ser significativo ao nível de 5%. A estimativa de $\alpha = 4,738$ resulta na variância da distribuição de fragilidade igual a $Var(V) = 1/\alpha = 0,211$, a qual é significativa e representa a heterogeneidade não observada de fatores externos. A estimativa negativa de β_1 indica que taxa de sobrevivência é menor para pacientes nos estágios III e IV.

Tabela 4.6: Resultados da análise dos dados considerando a covariável estágio do melanoma no modelo (3.11).

Parâmetro	EMV	IC (95%)		DP	*Wald*	P-Valor
		LI	LS			
λ	0,333	0,309	0,357	0,012	26,736	0,000
α	4,738	0,929	8,548	1,944	2,438	0,015
β_0	1,533	1,428	1,638	0,054	28,593	0,000
β_1	-3,044	-3,286	-2,802	0,123	-24,654	0,000

Na Figura 4.12 é possível notar que o modelo não se ajustou tão bem aos dados com a covariável referente ao estágio. Para os estágios I e II, p_0 estimado foi de 0,823, enquanto para os estágios III e IV foi de 0,181, indicando que os estágios mais avançados tem proporção de cura muito baixa.

- **Covariável Cirurgia**

  A análise da Tabela 4.7, revela a significância estatística dos parâmetros estimados com a inclusão da covariável cirurgia. Todos os parâmetros demonstram p-valores muito baixos, exceto pelo α, que tem um p-valor próximo de 1%, sendo assim há evidências da significância estatística de todos os parâmetros com a inclusão dessa covariável. A estimativa de $\alpha = 3,249$ resulta na variância da distribuição de fragilidade igual a $Var(V) = 1/\alpha = 0.308$, a qual é significativa e representa a heterogeneidade não observada de fatores externos. A estimativa positiva de β_1 indica que taxa de sobrevivência é maior para pacientes que realizaram a cirurgia referente ao melanoma.

  Na Figura 4.13 é possível notar que o modelo se ajustou melhor para pacientes

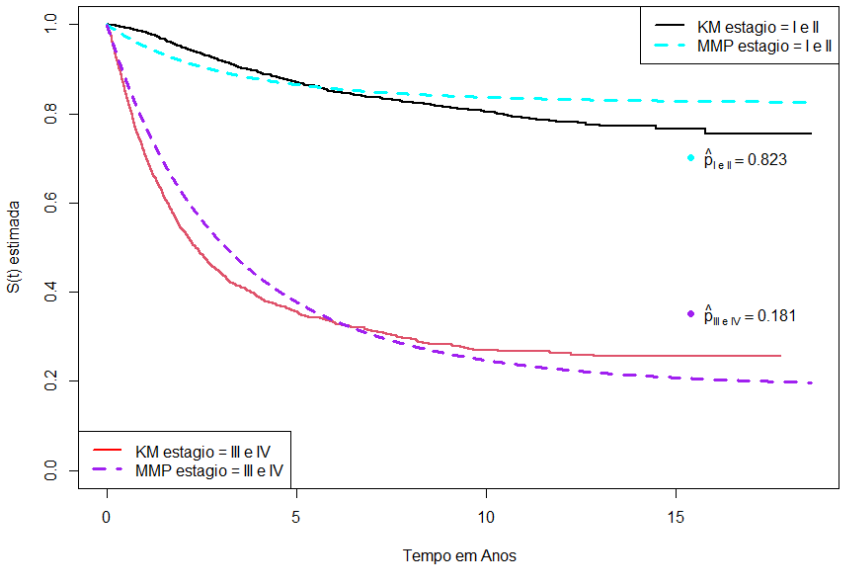

Figura 4.12: Curvas de sobrevivência para o modelo (3.11) com a inclusão da covariável estágio do melanoma.

Tabela 4.7: Resultados da análise dos dados considerando a covariável cirurgia no modelo (3.11).

Parâmetro	EMV	IC (95%)		DP	*Wald*	P-Valor
		LI	LS			
λ	0,308	0,281	0,334	0,013	23,277	0,000
α	3,249	0,812	5,686	1,243	2,613	0,009
β_0	-1,688	-2,186	-1,189	0,254	-6,638	0,000
β_1	2,261	1,836	2,686	0,217	10,437	0,000

que realizaram a cirurgia, pois a linha tracejada (em roxo) quase sobrepõe a curva de *Kaplan-Meier*. Para os o que não fizeram cirurgia o modelo não se ajustou tão bem e o p_0 estimado foi de 0,156, esse valor é bem baixo, pois conforme visto anteriormente, a cirurgia não é realizada em locais muito complicados ou estágios muito avançados. Para aqueles que fizeram a cirurgia, p_0 foi 0,64, indicando que pacientes que realizaram cirurgia tem proporção de cura mais alta.

- **Covariável Radioterapia**

A análise da Tabela 4.8, referente ao tratamento com radioterapia, revela a significância estatística dos parâmetros estimados com a inclusão dessa covariável. Todos os parâmetros exibem valores de p muito baixos, com exceção de α, que é próximo 1%, sugerindo a significância estatística dos parâmetros. A estimativa de $\alpha = 3,149$ resulta na variância da distribuição de fragilidade igual a $Var(V) = 1/\alpha = 0,318$,

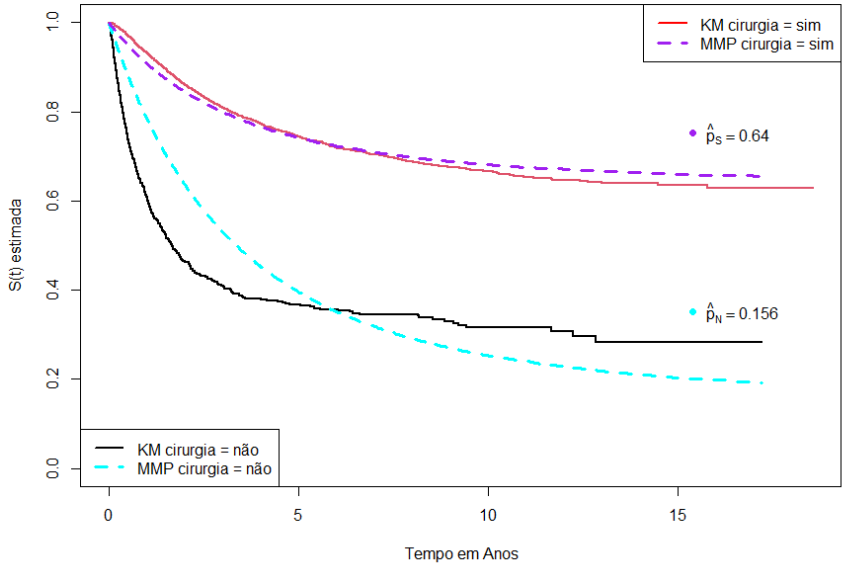

Figura 4.13: Curvas de sobrevivência para o modelo 4.7, considerando a variável cirurgia

a qual é significativa e representa a heterogeneidade não observada de fatores externos. A estimativa negativa de β_1 indica que taxa de sobrevivência é menor para pacientes que passaram pelo tratamento de radioterapia.

Tabela 4.8: Resultados da análise dos dados considerando a covariável radioterapia no modelo (3.11).

Parâmetro	EMV	IC (95%)		DP	*Wald*	P-Valor
		LI	LS			
λ	0,304	0,278	0,330	0,013	23,097	0,000
α	3,149	0,756	5,542	1,221	2,579	0,009
β_0	0,538	0,408	0,668	0,066	8,126	0,000
β_1	-3,218	-4,328	-2,108	0,566	-5,681	0,000

Na Figura 4.14, nota-se que o modelo se ajustou melhor para os pacientes que não realizaram o tratamento. Adicionalmente, o valor estimado de p_0 foi de 0,064 para os pacientes submetidos à radioterapia e de 0,631 para aqueles que não realizaram o tratamento.

- **Covariável Quimioterapia**

  A análise da Tabela 4.9, referente à quimioterapia, revela a significância estatística dos parâmetros estimados com a inclusão dessa covariável. Todos os parâmetros exibem valores de p muito baixos, com exceção de α que é 0,016, sugerindo a significância estatística dos parâmetros. A estimativa de $\alpha = 4,417$ resulta na variância da distribuição de fragilidade igual a $Var(V) = 1/\alpha = 0,226$, a qual é significativa

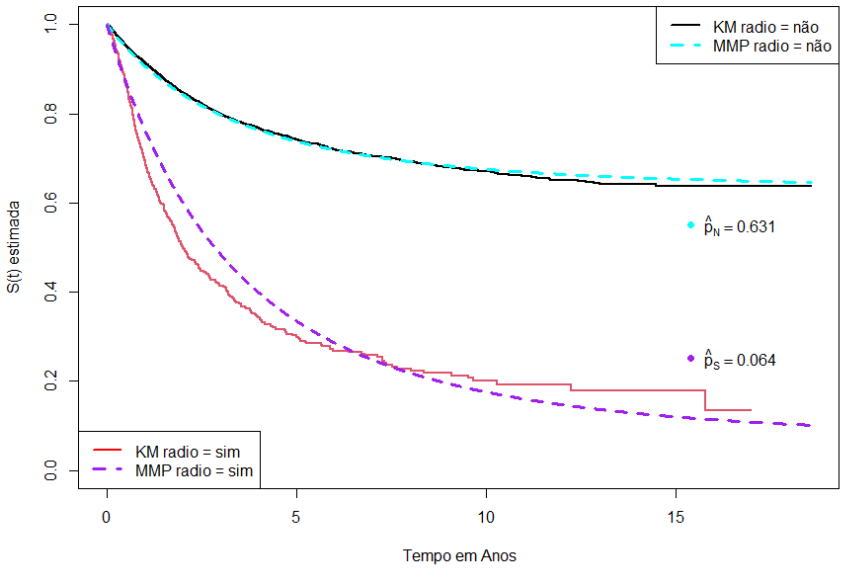

Figura 4.14: Curvas de sobrevivência para o modelo (3.11) com a inclusão da covariável radioterapia.

e representa a heterogeneidade não observada de fatores externos. A estimativa negativa de β_1 indica que taxa de sobrevivência é menor para pacientes que realizaram o tratamento de quimioterapia.

Tabela 4.9: Resultados da análise dos dados considerando a covariável quimioterapia no modelo (3.11).

Parâmetro	EMV	IC (95%)		DP	*Wald*	P-Valor
		LI	LS			
λ	0,324	0,299	0,348	0,013	25,848	0,000
α	4,417	0,831	8,002	1,829	2,414	0,016
β_0	0,856	0,756	0,955	0,051	16,851	0,000
β_1	-2,865	-3,268	-2,462	0,206	-13,934	0,000

A Figura 4.15 mostra que o modelo se ajustou um pouco melhor para os pacientes que não fizeram o tratamento. Adicionalmente, o valor estimado de p_0 foi de 0,118 para os pacientes submetidos à quimioterapia e de 0,702 para aqueles que não realizaram o tratamento.

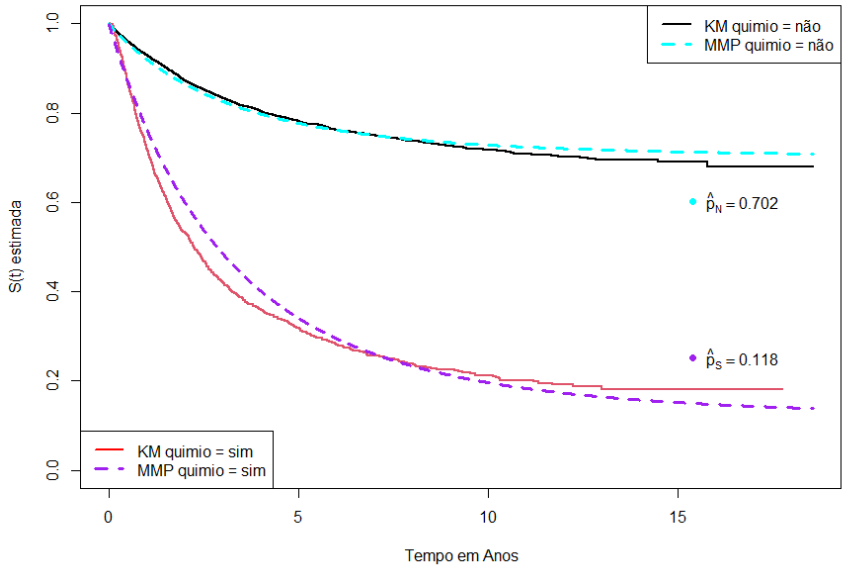

Figura 4.15: Curvas de sobrevivência para o modelo 3.11) com a inclusão da covariável quimioterapia.

Para as covariáveis radioterapia e quimioterapia, a proporção de cura e a taxa de sobrevivência foram melhores para os pacientes que não passaram por esses tratamentos. Esse resultado pode ser justificado pelo fato de que, na maioria dos casos em que os pacientes precisam de radioterapia e/ou quimioterapia, encontram-se em estágios mais avançados da doença, o que implica em uma probabilidade menor de sobrevivência.

Capítulo 5

Conclusão

Neste trabalho foi estudado o modelo de mistura padrão com fragilidade gama, que tem como principal característica a inclusão dos indivíduos que não estão suscetíveis ao evento de interesse, mesmo que o estudo tenha um longo período de tempo, além disso foi associada a existência de fatores externos não mensurados, ou seja, a fragilidade.

No início foram feitas revisões da análise de sobrevivência, explorando os tipos de modelagem em situações de longa duração e conceitos essenciais como os modelos de mistura padrão e fragilidade. Além disso, a explicação detalhada do modelo proposto, que combina o modelo mistura padrão com a inclusão da fragilidade gama, foi estruturada.

Foram feitas aplicações com dados do pacotes do *software R Studio*, em algumas seções, mas a aplicação prática do modelo em estudo foi aplicada aos dados reais de melanoma, obtidos pela Fundação Oncocentro de São Paulo (FOSP). A análise dos dados resultou em *insights* relevantes como, por exemplo, a fração de cura de 60% dos pacientes, além disso a diferença da taxa de sobrevivência para diferentes grupos. Os pacientes do sexo feminino tem uma taxa de sobrevivência mais alta, assim como os pacientes que realizaram a cirurgia relacionada ao melanoma e o pacientes nos estágios I e II, porém a realização de tratamentos (radioterapia e quimioterapia) tem uma taxa de sobrevivência mais baixa, isso porque esses tratamentos são utilizados em estágios mais avançados. Os resultados obtidos destacam a eficácia do modelo de mistura padrão com fragilidade gama tanto sem covariáveis quanto com a inclusão delas, cada covariável incluída individualmente foi estatisticamente significativa.

A aplicação deste modelo pode se beneficiar de estudos futuros que explorem ainda mais suas capacidades em diferentes conjuntos de dados e contextos clínicos e contribuir para o avanço do conhecimento em análise de sobrevivência para modelagem em situações

de longa duração. Considerando propostas futuras, é possível realizar a estimação através de outros métodos, além da análise mais profunda do modelo e análise de resíduos.

Referências Bibliográficas

American Cancer Society (2023). *Melanoma Skin Cancer: Treating Melanoma by Stage*.

Anand, P., Kunnumakara, A., Sundaram, C. e et al. (2008). Cancer is a preventable disease that requires major lifestyle changes. *Pharm Res*, **25**, 2097–2116.

Azevedo Silva, R. d. (2011). *Modelos de fração de cura com fatores latentes competitivos e fragilidade*. Dissertação de mestrado em estatística, Universidade de São Paulo, São Paulo. [acesso 2023-08-21].

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.

Bolfarine, H. e Sandoval, M. C. (2001). *Introdução à Inferência Estatística, Volume 2*. SBM.

Brown, B. e Flood, M. (1947). Tumbler mortality. *Journal of the American Statistical Association*, **42**, 562–574.

Calsavara, V. F., Milani, E. A., Bertolli, E. e Tomazella, V. (2020). Long-term frailty modeling using a non-proportional hazards model: Application with a melanoma dataset. *Statistical Methods in Medical Research*, **29**(8), 2100–2118.

Chen, M., Ibrahim, J. 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.

Colosimo, E. A. e Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*. Number 15 Em

Série do livro. Edgard Blucher, São Paulo. ISBN XX-XXX-XXXX-X. Bibliografia: p. 131–132.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2), 187–202.

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

Elbers, C. e Ridder, G. (1982). True and spurius duration dependence: (t)he identifiabiability of the proportional hazard model. *Journal of the Royal Statistical Society: Series B (Methodological)*, páginas 403–409.

Espirito Santo, A. P. J. d. (2022). *Modelos de sobrevivência induzidos por fragilidade discreta com fração de cura e riscos proporcionais*. Tese de doutorado em estatística, Universidade de São Paulo, São Carlos. [acesso 2023-08-21].

Ibrahim, J. G., Chen, M.-H. e Sinha, D. (2014). B ayesian survival analysis. *Wiley StatsRef: Statistics Reference Online*.

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

Kirkwood, T. e Austad, S. (2000). Why do we age? *Nature*, **408**, 233–238.

Klein, J. P. e Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, second edition.

Kuk, A. e Chen, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biomatrika*, **79**, 531–541.

Laurie, J., Moertel, C., Fleming, T., Wieand, H., Leigh, J., Rubin, J., McCormack, G., Gerstner, J., Krook, J. 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. *J Clinical Oncology*, **7**, 1447–1456.

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

Lee, E. (1980). *Statistical Methods for Survival Data Analysis*. Lifetime Learning Publications, New York.

Molina, K., Calsavara, V. F., Tomazella, V. D. e Milani, E. A. (2021). Survival models induced by zero-modified power series discrete frailty: Application with a melanoma data set. *Statistical Methods in Medical Research*, **30**(8), 1874–1889.

Peng, Y. e Zhang, J. (2008). Estimation method of the semiparametric mixture cure gamma frailty model. *Statistcs i*, **27**, 5177–5194.

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.

Sociedade Brasileira de Cirurgia Oncológica (2023). *Câncer de Pele: Melanoma Maligno*.

Society, A. C. (2023). *Key Statistics for Melanoma Skin Cancer*.

Stacy, E. W. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, **33**(3), 1187–1192.

Taconeli, J. P. (2013). *Modelo de mistura paramétrico com fragilidade na presença de covariáveis*. Dissertação (mestrado em ciências exatas e da terra), Universidade Federal de São Carlos, São Carlos. 75 f.

Tomazella, V. d. L. D. (2003). *Modelagem de dados de eventos recorrentes via processos de Poisson com termo de fragilidade*. Doutorado em ciências de computação e matemática computacional, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

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.

Vera Tomazella, P. D. (2022). *Aulas de Análise de Sobrevivência*. Material do curso.

Weibull, W. (1939). A statistical theory of the strength of materials, 1939. *Ingeniors Vetenskaps Akademien Handlingar*, (n.151: The Phenomenon of Rupture in Solid), 292–297.

Wienke, A. (2007). *Frailty models in survival analysis*. Chapman and Hall/CRC.

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