



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Câmpus de Presidente Prudente

Ana Clara Rodrigues da Silva

**AVALIAÇÃO DE PACIENTES COM CÂNCER DE MAMA: UMA ABORDAGEM
UTILIZANDO MODELOS DE FRAÇÃO DE CURA WEIBULL**

PRESIDENTE PRUDENTE

2025

Ana Clara Rodrigues da Silva

**AVALIAÇÃO DE PACIENTES COM CÂNCER DE MAMA: UMA ABORDAGEM
UTILIZANDO MODELOS DE FRAÇÃO DE CURA WEIBULL**

Trabalho de Conclusão de Curso apresentado
ao Curso de Graduação em Estatística da
FCT/Unesp para aproveitamento na disciplina
Trabalho de Conclusão de Curso.

Orientador: Prof. Dr. Ricardo Puziol de
Oliveira

PRESIDENTE PRUDENTE

2025

S586a Silva, Ana Clara Rodrigues da
Avaliação de pacientes com câncer de mama: Uma abordagem utilizando modelos de fração de cura Weibull / Ana Clara Rodrigues da Silva. -- Presidente Prudente, 2025

41 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Ricardo Puziol de Oliveira

1. Distribuição Weibull. 2. Fração de cura de mistura. 3. Censura. 4. Curva de sobrevivência. 5. Resíduos de Cox-Snell.
I. Título.

TERMO DE APROVAÇÃO

ANA CLARA RODRIGUES DA SILVA

AVALIAÇÃO DE PACIENTES COM CÂNCER DE MAMA: UMA ABORDAGEM UTILIZANDO MODELOS DE FRAÇÃO DE CURA WEIBULL

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador:



Documento assinado digitalmente
RICARDO PUZIOL DE OLIVEIRA
Data: 01/12/2025 22:47:41-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Ricardo Puziol de Oliveira

Departamento de Estatística



Documento assinado digitalmente
FERNANDO ANTONIO MOALA
Data: 02/12/2025 22:34:34-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Fernando Antônio Moala

Departamento de Estatística

Presidente Prudente, 28 de novembro de 2025.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus e à Nossa Senhora, por me guiarem, permitirem chegar até aqui, alcançando o primeiro de muitos objetivos que ainda virão.

Aos meus pais, Luiz Vanderlei e Natalina, por todo amor, dedicação e apoio incondicional, e por acreditarem em mim mesmo quando eu mesma duvidava.

Aos meus irmãos, Luiz Gabriel e Larissa, fonte de inspiração diária. Obrigada por me ouvirem nas minhas angústias, por me animarem nos dias difíceis, por dividirem comigo alegrias e frustrações e, sobretudo, por nunca me deixarem desistir. Cada palavra de incentivo, cada gesto de carinho foi essencial para que eu chegasse até aqui.

Aos meus amigos, que estiveram comigo em cada fase, nos choros, nas risadas. Obrigada por me ouvirem, me entenderem e me lembrarem, com tanto carinho, que eu era capaz.

Ao meu orientador, professor Ricardo, pela paciência, ensinamentos e por me conduzir com sabedoria e dedicação durante todo o processo.

E a todos os professores do curso de Estatística da UNESP de Presidente Prudente, que contribuíram de forma valiosa para minha formação, transmitindo conhecimento e inspirando amor pela profissão.

A cada um de vocês, meu sincero e profundo agradecimento. Sem o apoio, o carinho e a presença de todos, este sonho não teria se tornado possível.

RESUMO

O trabalho tem como objetivo analisar os fatores associados ao câncer de mama, por meio da aplicação de um modelo de fração de cura do tipo com mistura baseado na distribuição weibull, voltado para estudos de sobrevivência em que parte da população é imune ao evento de interesse. Para o desenvolvimento do presente estudo, foi utilizada a base de dados NKI Breast Cancer, proveniente do Netherlands Cancer Institute, obtida de forma secundária a partir do estudo de Van de Vijver et al. (2002). A base é composta por informações de 272 mulheres diagnosticadas com câncer de mama primário invasivo, cujos tumores apresentavam diâmetro inferior a 5 cm (classificação pT1 ou pT2), com diagnósticos realizados entre os anos de 1984 e 1995. O modelo mostrou-se adequado para capturar a heterogeneidade dos dados e descrever de forma eficiente situações em que há indivíduos curados, contribuindo para o avanço de metodologias aplicáveis em análises de sobrevivência multivariadas.

Palavras-chave: Análise de sobrevivência; Modelo Weibull; Fração de cura; Kaplan-Meier.

ABSTRACT

This study aims to analyze the factors associated with breast cancer through the application of a mixture cure rate model based on the Weibull distribution, focused on survival studies in which part of the population is immune to the event of interest. For the development of this research, the NKI Breast Cancer dataset from the Netherlands Cancer Institute was used, obtained secondarily from the study by Van de Vijver et al. (2002). The dataset consists of information from 272 women diagnosed with primary invasive breast cancer, whose tumors had a diameter smaller than 5 cm (classified as pT1 or pT2), with diagnoses made between 1984 and 1995. The model proved to be suitable for capturing data heterogeneity and efficiently describing situations in which there are cured individuals, contributing to the advancement of methodologies applicable to multivariate survival analysis.

Keywords: Survival analysis; Weibull model; Cure rate; Kaplan-Meier.

LISTA DE FIGURAS

Figura 1. Tipos de censura em estudos de sobrevivência.....	19
Figura 2. Curva de sobrevivência estimada via Kaplan-Meier referente ao tempo de sobrevida dos pacientes com câncer de mama.....	30
Figura 3. Matriz de correlação de Spearman entre as covariáveis consideradas no estudo.....	32
Figura 4. Curva de sobrevivência empírica (estimada via Kaplan-Meier) vs curva de sobrevivência ajustada pelo modelo de fração de cura de mistura Weibull para o tempo de sobrevivência dos pacientes com câncer de mama.....	33
Figura 5. Resíduos de Cox-Snell para o modelo de fração de cura de mistura Weibull proposto.....	35

LISTA DE TABELAS

Tabela 1. Características das 272 pacientes com câncer de mama do Instituto do Câncer da Holanda do estudo de Van de Vijver et al. (2002).....	23
Tabela 2. Resultado do estudo de simulação Monte Carlo: $\lambda = 2.0$, $\gamma = 1.5$ e $\rho = 0.3$	29
Tabela 3. Resultado do estudo de simulação Monte Carlo: $\lambda = 1.5$, $\gamma = 1.0$ e $\rho = 0.1$	29
Tabela 4. Resultado do estudo de simulação Monte Carlo: $\lambda = 2.5$, $\gamma = 2.5$ e $\rho = 0.5$	30
Tabela 5. Estimativas de parâmetros e intervalos de confiança para 95% do modelo de regressão de fração de cura proposto.....	34

Sumário

1 INTRODUÇÃO.....	11
2 REFERENCIAL TEÓRICO	14
2.1. TRABALHOS RELACIONADOS	14
2.2. CONCEITOS DE ANÁLISE DE SOBREVIVÊNCIA	16
2.3. MECANISMOS DE CENSURA	18
2.4. ESTIMADOR KAPLAN-MEIER.....	20
2.5. MODELOS DE FRAÇÃO DE CURA.....	21
3 MATERIAIS E MÉTODOS.....	22
3.1. BASE DE DADOS	22
3.2. DISTRIBUIÇÃO WEIBULL	23
3.2.1. INCORPORAÇÃO DA FRAÇÃO DE CURA	24
3.2.2. INCORPORAÇÃO DE COVARIÁVEIS	26
3.3. RESÍDUOS DE COX-SNELL	26
4 ESTUDO DE SIMULAÇÃO MONTE CARLO	28
5. RESULTADOS E DISCUSSÕES	30
5.1. ANÁLISE EXPLORATORIA DE DADOS	30
5.2. ANÁLISE PELO MODELO DE FRAÇÃO DE CURA WEIBULL	32
6. CONCLUSÃO	36
REFERÊNCIAS	37

1 INTRODUÇÃO

O câncer de mama (CM), de forma geral, é uma condição crônica que surge a partir de sucessivas mutações no material genético das células. Essas mutações resultam em mudanças no ciclo celular, levando a uma proliferação desregulada e à falha no processo de apoptose, o que causa o crescimento incontrolável de células anormais que podem invadir tecidos vizinhos e se disseminar pelo corpo, sendo os principais fatores prognósticos, em geral, a idade, o tamanho do tumor, o *status* dos linfonodos axilares, o tipo histológico do tumor, o grau patológico e o status de receptores hormonais. (HASSANPOUR; DEHGHANI, 2017; INCA, 2022).

No ano de 2020, as estimativas do *Global Cancer Observatory (Globocan)*, da *International Agency for Research on Cancer (IARC)*, e da Organização Mundial da Saúde (OMS) indicaram aproximadamente 19,3 milhões de novos casos de câncer e quase 10 milhões de mortes em todo o mundo (INCA, 2022; OMS, 2022; SUNG, 2021). Em 2022, nos dados publicados pela Agência Internacional de Pesquisa sobre o Câncer (IARC) e apresentadas no Câncer Today do Observatório Global do Câncer (FERLAY, 2024) foram também registrados aproximadamente 20 milhões de novos casos de câncer e 9,7 milhões de mortes relacionadas à doença. Entre as mulheres, o tipo de câncer mais diagnosticado é o CM, com uma incidência de 11.6% de novos casos. Além disso, o CM é o quarto câncer com maior número de mortes, apresentando uma taxa de 6.9% de eventos. Esses dados refletem, em parte, a disparidade global no acesso ao diagnóstico precoce e ao tratamento adequado: embora avanços significativos nos métodos de rastreamento e diagnóstico tenham contribuído para a redução da mortalidade em regiões de maior renda, países de baixa e média renda ainda enfrentam o desafio de diagnósticos tardios, frequentemente em estágios avançados da doença, quando as possibilidades terapêuticas são mais restritas e menos eficazes (BRAY, 2018; UNGER-SALDAÑA, 2014).

Neste cenário, o presente estudo visa analisar os fatores associados ao câncer de mama, com foco na identificação de variáveis que influenciam a sobrevivência dos pacientes. Como há a possibilidade de remissão completa em uma parcela dos casos, que implica na presença de um *platô* na curva de sobrevivência — que sugere a existência de uma fração da população que não está mais sujeita ao risco de falha

(óbito), isto é, pacientes considerados sobreviventes de longo prazo —, a abordagem estatística adotada deve levar em conta essa heterogeneidade. Sendo assim, este estudo considera uma abordagem de análise por meio de modelos de fração de cura, usando a distribuição Weibull como *baseline* para a construção do modelo. A escolha dessa distribuição justifica-se por sua reconhecida aplicabilidade em estudos de sobrevivência, destacando-se pela flexibilidade na adaptação a diferentes perfis de dados e pela simplicidade na obtenção de inferências significativas.

A motivação deste estudo reside no contínuo avanço de técnicas estatísticas, particularmente aquelas voltadas à análise de sobrevivência, para compreender as complexidades do câncer de mama e os múltiplos fatores que influenciam a sobrevida das pacientes. Moore (2016), por exemplo, utilizou modelos de sobrevivência para avaliar a eficácia de diferentes tratamentos em pacientes com câncer de mama em estágio avançado, destacando como variáveis clínicas, como a idade e a presença de mutações genéticas, afetam a taxa de resposta ao tratamento. Adicionalmente, estudos publicados no *PubMed* expandiram o entendimento sobre o câncer de mama, abordando fatores genéticos, avanços terapêuticos e técnicas de análise de sobrevivência. Mutações nos genes BRCA1 e BRCA2, que ocorrem em 5-10% dos casos hereditários, aumentam o risco de recorrência e reduzem a sobrevida sem tratamento adequado. Terapias como inibidores de PARP mostraram-se eficazes para pacientes com essas mutações. Além disso, a combinação de quimioterapia e tratamentos hormonais tem prolongado a sobrevida em casos com receptor hormonal positivo, destacando a importância de pesquisas contínuas para aprimorar diagnósticos e tratamentos.

Para além dos fatores genéticos e terapêuticos, aspectos comportamentais e antecedentes familiares também desempenham papel relevante no risco de desenvolvimento da doença. O histórico familiar aumenta de duas a três vezes a probabilidade de ocorrência de câncer de mama; no entanto, apenas uma em cada dez mulheres diagnosticadas possui casos registrados em familiares (INCA, 2013b). Em contrapartida, fatores modificáveis como alimentação saudável e prática regular de atividade física contribuem para a redução do risco, sendo estimado que aproximadamente 30% dos casos poderiam ser evitados por meio da adoção de tais medidas (INCA, 2013b).

Além de sua relevância estatística e clínica, este estudo também se alinha com o Objetivo de Desenvolvimento Sustentável (ODS) número 3 da Organização das

Nações Unidas (ONU), que visa garantir uma vida saudável e promover o bem-estar para todas as pessoas em todas as idades. No caso, a pesquisa em câncer de mama e o aprimoramento das técnicas de análise de sobrevivência contribuem diretamente para as metas relacionadas à prevenção, diagnóstico precoce e tratamento eficaz de doenças, reforçando a importância dessa investigação científica no alcance dessas metas globais.

OBJETIVOS

OBJETIVO GERAL

Conforme explícito anteriormente, este estudo tem por objetivo geral analisar os fatores associados ao câncer de mama, com foco na identificação das variáveis que influenciam a sobrevivência dos pacientes. Para tanto, será considerada uma abordagem de análise baseada em modelos de fração de cura de mistura, usando a distribuição Weibull como *baseline*, visando a incorporação da presença de um platô na curva de sobrevivência que está diretamente ligada à possibilidade de remissão completa em uma parcela dos pacientes.

OBJETIVOS ESPECÍFICOS

- Fazer uma revisão de literatura sobre modelos de fração de cura, com ênfase nas abordagens de mistura e no uso da distribuição Weibull em contextos clínicos, particularmente, oncológicos.
- Fazer uma análise exploratória dos dados, a fim de identificar quais dos fatores clínicos e laboratoriais estão mais relacionados com a sobrevida dos pacientes com câncer de mama.
- Construir um modelo de fração de cura usando abordagem de mistura baseada na distribuição Weibull, além de incorporar uma estrutura de regressão no parâmetro de cura, ρ , para identificar quais dos fatores estão associados a probabilidade de “cura”, isto é, a probabilidade de remissão do câncer.
- Validação do modelo proposto utilizando medidas de desempenho como, por exemplo, os critérios de informação, e, também, diagnósticos de resíduos, com

base nos resíduos de Cox-Snell modificados para o contexto de modelos de fração de cura.

2 REFERENCIAL TEÓRICO

2.1. TRABALHOS RELACIONADOS

O câncer de mama é uma neoplasia maligna que se origina nas células dos ductos ou lóbulos do tecido mamário. O desenvolvimento do câncer de mama ocorre devido a mutações genéticas que desregulam o ciclo celular, levando à proliferação descontrolada de células anormais que podem invadir tecidos adjacentes e formar metástases (HASSANPOUR & DEHGHANI, 2017). Durante a iniciação, mutações genéticas acumulam-se nas células, e, na promoção, fatores internos e externos contribuem para a proliferação celular desregulada. A progressão ocorre quando as células tumorais se tornam malignas, ganhando a capacidade de invadir outros tecidos e formar metástases (WEINBERG, 2013).

Diversos fatores de risco estão associados ao desenvolvimento do câncer de mama, como idade avançada, histórico familiar e mutações nos genes BRCA1 e BRCA2, que aumentam significativamente o risco da doença e influenciam o prognóstico e as opções de tratamento (FOULKES, 2008). Fatores hormonais, como o uso de terapia de reposição hormonal, além de hábitos de vida como sedentarismo e obesidade, também são conhecidos por elevar o risco de câncer de mama e podem impactar a sobrevida e a eficácia dos tratamentos (GRUPO COLABORATIVO SOBRE FATORES HORMONAIS NO CÂNCER DE MAMA, 2012). Os tratamentos para câncer de mama variam conforme o estágio da doença e as características do tumor, incluindo cirurgia (mastectomia ou conservadora), quimioterapia, radioterapia e terapias hormonais, cada uma com impacto específico nos desfechos clínicos (JUN WANG, 2019). Terapias direcionadas, como os inibidores de HER2, têm demonstrado avanços importantes na sobrevida de pacientes com câncer de mama metastático (CARDOSO, 2020).

Os modelos de fração de cura têm sido amplamente explorados na análise de sobrevivência para descrever a ocorrência de eventos de interesse, como recidiva e cura. No estudo de Gallardo et al. (2024), por exemplo, foi desenvolvido um novo modelo de sobrevivência com fração de cura que considera causas competidoras, em que o número de causas segue uma mistura entre as distribuições de Poisson e

Birnbaum-Saunders. A proposta metodológica, estimada via máxima verossimilhança com o algoritmo EM e avaliada por simulações de Monte Carlo, foi aplicada a um conjunto de dados de câncer de mama provenientes de pacientes atendidas em São Paulo, Brasil, permitindo analisar simultaneamente os efeitos de covariáveis sobre o número de causas competidoras e sobre a taxa de cura. Os resultados mostraram que o modelo apresentou desempenho superior em relação a alternativas clássicas, fornecendo estimativas mais realistas para a probabilidade de cura e para o risco residual, o que reforça sua utilidade em estudos clínicos e epidemiológicos baseados em dados reais.

No estudo de Lima et al. (2025), por exemplo, foram analisados dados nacionais de mulheres com câncer de mama utilizando modelos aditivos generalizados para localização, escala e forma (GAMLSS), com ênfase na identificação de fatores associados à sobrevida. A base de dados contemplou informações clínicas e sociodemográficas, e os resultados mostraram que variáveis como estágio da doença, grau histológico, idade ao diagnóstico e região de residência apresentaram efeitos estatisticamente significativos sobre o risco de mortalidade. O uso do GAMLSS permitiu avaliar não apenas a média, mas também a dispersão e a forma da distribuição dos tempos de sobrevida, o que evidenciou a heterogeneidade existente entre os diferentes perfis de pacientes. Esse enfoque destacou a relevância do estágio avançado e de piores características histológicas como determinantes de menor sobrevida, enquanto idade mais jovem e residir em determinadas regiões do país estiveram associadas a prognósticos diferenciados, ressaltando a importância de análises detalhadas de dados nacionais para subsidiar políticas de saúde no contexto do câncer de mama no Brasil.

No estudo de Rodrigues (2025), por exemplo, foi desenvolvido um modelo de regressão quantílica para dados de sobrevivência com presença de sobreviventes a longo prazo, baseado na distribuição de Gompertz defectiva. A proposta metodológica permite estimar quantis específicos da distribuição de sobrevivência, considerando a presença de uma fração de indivíduos curados, o que é comum em estudos oncológicos. O modelo foi aplicado a dados de câncer de próstata em homens, utilizando informações clínicas e demográficas, como idade, estágio da doença e histórico familiar. Os resultados indicam que variáveis como idade ao diagnóstico e estágio clínico mostraram-se estatisticamente significativas na determinação dos quantis de sobrevivência, evidenciando a utilidade do modelo para fornecer

estimativas mais precisas e específicas do tempo de sobrevida em diferentes percentis da população.

No estudo de Gordon (1996), por outro lado, foi aplicado um modelo de mistura de cura para analisar dados de sobrevida de pacientes com câncer de mama, distinguindo aqueles que morreriam por causas relacionadas ao câncer dos que faleceriam por outras causas. Utilizando dados censitários e de ensaios clínicos, estimou-se que 55,9% das pacientes com câncer de mama sem linfonodos comprometidos (node negative) e 26,0% das com linfonodos comprometidos (node positive) morreriam por causas não relacionadas ao câncer. Entre as pacientes que faleceriam devido ao câncer de mama, as do grupo node positive apresentaram uma taxa de mortalidade significativamente mais alta, indicando que essa forma de câncer não seria apenas uma manifestação tardia da doença node negative, mas sim uma condição mais agressiva.

Por fim, no estudo de Shabankhani et al. (2022), foi realizada uma análise de sobrevida de cinco anos em pacientes com câncer de mama na província de Mazandaran, Irã, utilizando modelos de cura. A pesquisa envolveu 221 mulheres diagnosticadas entre 2014 e 2017, acompanhadas até 2021, e os dados foram analisados com softwares SPSS 23 e R 3.6.3. Os resultados indicaram que 45% das mortes ocorreram após o tratamento, com taxas de cura de 97%, 84% e 75% após um, três e cinco anos, respectivamente. O modelo de Weibull não misturado apresentou o melhor ajuste ($AIC=659$), estimando uma taxa de cura de 68%. Variáveis como idade ao diagnóstico, tabagismo, envolvimento de linfonodos, quimioterapia e necessidade de hospitalização de emergência mostraram-se estatisticamente significativas, com a presença de linfonodos comprometidos associada a uma redução de 85% na taxa de cura, e a quimioterapia a uma redução de 12%. Esses achados reforçam a importância de modelos de cura na análise de dados de longo prazo em câncer de mama, permitindo estimativas mais precisas da sobrevida e identificação de fatores prognósticos críticos.

2.2. CONCEITOS DE ANÁLISE DE SOBREVIVÊNCIA

O estudo do **tempo até que um evento ocorra**, frequentemente denominado dados de sobrevivência, tempo de vida ou tempo até a falha, envolve a análise do intervalo necessário para que um acontecimento específico se realize. Esse tipo de

análise é utilizado em diferentes áreas, como medicina, biologia, saúde pública, epidemiologia, engenharia, economia e demografia. Conforme Tutz et al. (2016), seu principal propósito é compreender a duração até o evento, possibilitando estimar probabilidades de sobrevivência ou resposta, calcular tempos médios de vida, comparar distribuições e identificar fatores de risco ou variáveis capazes de influenciar esse tempo.

O conceito de “evento” é flexível e depende do contexto: na pesquisa médica, pode incluir a morte, a cura de uma doença, a remissão de um câncer específico, a infecção de indivíduos expostos a um patógeno ou ainda tempos associados a tratamentos em órgãos ou tecidos humanos; na engenharia, o evento pode ser a falha, reparo ou desgaste de componentes de máquinas; em toxicologia, pode referir-se à sobrevivência de filhotes em cada ninhada ou ao efeito de diferentes drogas na atividade cerebral. Outras aplicações incluem o tempo até liberdade condicional de condenados, a duração de casamentos, o tempo até a assinatura de jornais ou revistas, ou ainda pedidos de indenização trabalhista, sempre considerando fatores de risco ou previsões relacionados ao tempo até o evento (COX, 1972; MALLER AND ZHOU, 1996; KLEIN AND MOESCHBERGER, 1997; COLLETT, 2015).

Considere um paciente selecionado aleatoriamente da população, cujo tempo até a ocorrência do evento de interesse seja representado por T . Conforme destacado por Li e Ma (2013), é fundamental ajustar uma distribuição de probabilidade adequada para a variável T , de modo a possibilitar respostas a perguntas práticas, como “por quanto tempo um paciente com câncer de pulmão pode sobreviver com 90% de probabilidade?”. Para tratar essas questões, duas funções desempenham papel central na análise de dados de sobrevivência: a função de sobrevivência e a função de risco. A função de sobrevivência representa a probabilidade de que o tempo até o evento exceda um determinado valor t , permitindo quantificar a chance de um paciente permanecer vivo além de um período específico. Em termos matemáticos, essa função é descrita por:

$$S(t) = P(T \geq t) = 1 - F(t) \quad (2.1)$$

onde $F(t)$, denotado como função de distribuição acumulada, é a probabilidade de um indivíduo selecionado aleatoriamente morrer antes do tempo t . Segundo Lawless

(1982), a função de sobrevivência não é crescente e contínua à esquerda, isto é, $S(0) = 1$ e $S(\infty) = 0$.

Outra função fundamental na análise de dados de sobrevivência é a função de risco, que descreve a probabilidade de ocorrência do evento em um instante específico t , considerando que o indivíduo tenha sobrevivido até esse ponto. Em outras palavras, a função de risco indica a chance de morrer exatamente no tempo t , dado que o paciente permaneceu vivo até então, fornecendo informações complementares à função de sobrevivência e permitindo uma compreensão mais detalhada do perfil temporal do risco associado ao evento. Matematicamente, tem-se a expressão:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid \geq t)}{\Delta t} \quad (2.2)$$

onde Δt é um *incremento infinitesimal de tempo*. Em algumas aplicações práticas, $h(t)$ também é chamada de *taxa instantânea de falha* e pode ser expressa como $h(t) = f(t)/S(t)$, onde $f(t)$ é a *função de densidade de probabilidade* (ou de massa, no caso discreto).

2.3. MECANISMOS DE CENSURA

De acordo com Klein e Moeschberger (1997), Giolo e Colosimo (2006) e Collett (2015), uma característica central dos dados de sobrevivência é a presença de observações censuradas. O tempo de sobrevivência de um indivíduo é considerado censurado quando o evento de interesse não é observado, geralmente porque o acompanhamento foi interrompido ou porque o paciente faleceu por uma causa não relacionada ao estudo. Dois tipos de censura são frequentemente considerados na literatura:

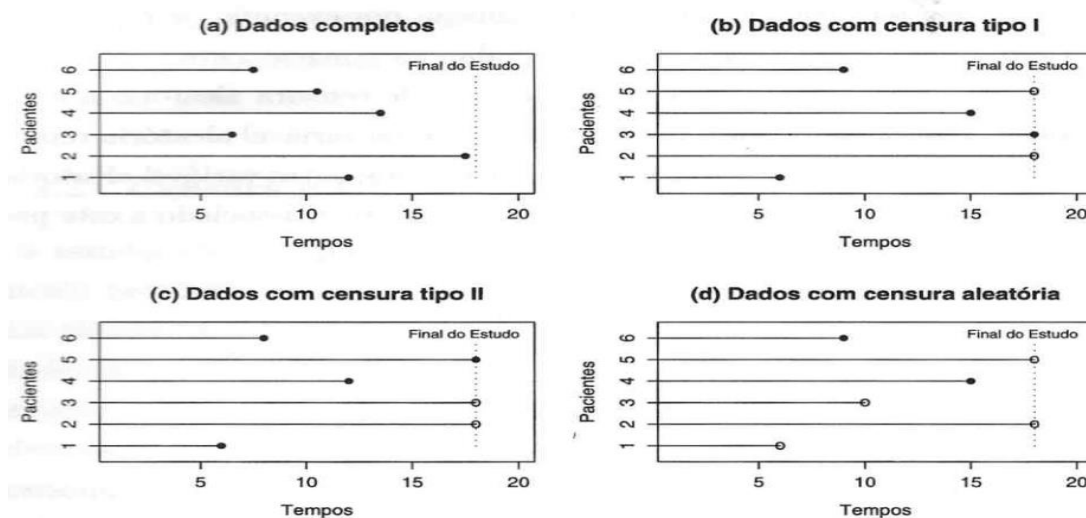
- **Censura Tipo I:** ocorre quando o estudo possui um período de acompanhamento pré-determinado. Nesse caso, os indivíduos que desenvolvem o evento de interesse durante o período são observações completas, enquanto aqueles que permanecem sem o evento até o final do estudo têm seus tempos de sobrevivência registrados apenas como pelo

menos iguais à duração total do acompanhamento. O tempo de censura é, portanto, igual ao comprimento do estudo, o que simplifica a estimação de parâmetros.

- **Censura Tipo II:** é mais comum em estudos laboratoriais com animais, em que o evento de interesse é geralmente fatal. O estudo continua até que um número ou proporção específica de indivíduos tenha apresentado o evento. Nesse caso, o tempo de censura corresponde ao maior tempo de sobrevivência observado entre os indivíduos não censurados.

Além destes dois tipos, ainda é possível a presença de censura aleatória que ocorre quando o tempo de censura varia entre os indivíduos, sem seguir um ponto fixo ou critério predeterminado. Esse tipo de censura é comum em estudos clínicos ou observacionais, em que os pacientes podem ser perdidos de acompanhamento em momentos distintos por motivos diversos. A Figura 1 apresenta um resumo visual dos diferentes tipos de censura, facilitando a compreensão das distinções entre eles.

Figura 1. Tipos de censura em estudos de sobrevivência.



Fonte: Análise de Sobrevivência, Giolo e Colosimo (2006).

O tipo de censura mais comum em estudos de sobrevivência é a censura à direita, geralmente resultante de limitações de tempo ou recursos, que impedem acompanhar todos os indivíduos até que o evento de interesse ocorra. Em termos formais, esse esquema de censura pode ser descrito por duas variáveis: T , que representa o tempo até a falha ou ocorrência do evento, e C , que indica o tempo em

que a observação é interrompida, sendo independente de T . A partir dessas variáveis, define-se o tempo efetivamente observado e o indicador de censura, possibilitando uma modelagem adequada dos dados censurados e permitindo estimativas corretas das funções de sobrevivência. Isto é,

$$t = \min(T, C) \quad \delta = \begin{cases} 1, & T \leq C \\ 0, & T > C \end{cases} \quad (2.3)$$

Por outro lado, o esquema de censura à esquerda também pode ser definido usando duas variáveis avaliadas, como no esquema de censura à direita. Nesse caso, t é igual a T se o tempo de vida para coletado e δ indica se o tempo exato de vida T foi coletado ($\delta = 1$) ou não ($\delta = 0$). Portanto, para a censura à esquerda, os dados coletados são dados por:

$$t = \max(T, C) \quad \delta = \begin{cases} 1, & T \geq C \\ 0, & T < C \end{cases} \quad (2.4)$$

Vale destacar que, se a presença de observações censuradas não for considerada, técnicas tradicionais de análise de regressão, geralmente baseadas em erros normais e dados transformados, poderiam ser aplicadas aos dados de sobrevivência. No entanto, essas abordagens — que assumem normalidade ou se apoiam em métodos não paramétricos — tornam-se inadequadas quando há censura ou presença de taxas de cura. Para lidar com essas situações, a literatura tem desenvolvido técnicas específicas, incluindo o modelo de riscos proporcionais de Cox (Cox, 1972), os modelos de longo prazo ou de taxa de cura (WIENKE, 2006) e os modelos de fragilidade (PRICE E MANATUNGA, 2001).

2.4. ESTIMADOR KAPLAN-MEIER

Em estudos de sobrevida, é comum observar que a curva de sobrevivência atinge um platô após certo período de acompanhamento. Segundo Lambert et al. (2007), isso ocorre quando a mortalidade dos indivíduos afetados se iguala à esperada na população geral, indicando que o excesso de risco deixa de existir. Esse padrão sugere a presença de uma fração de pacientes efetivamente curados, ou seja, que não apresentam mais risco associado à doença. Para identificar essa fração,

utilizam-se abordagens analíticas específicas, sendo a análise não-paramétrica de Kaplan–Meier um passo inicial importante para estimar a função de sobrevivência desses pacientes.

Para definir o estimador de Kaplan-Meier, considere $t_1 < \dots < t_m$ os instantes observados de ocorrência do evento. Para cada tempo t_j , define-se d_j como número de óbitos observados nesse instante e r_j como o número de pacientes em risco imediatamente antes de t_j . O *estimador de Kaplan-Meier* para a função de sobrevivência, denotado por $\hat{S}(t)$, é dado por:

$$\hat{S}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right) \quad (2.5)$$

Esse método produz uma curva de sobrevivência que descreve, de maneira empírica, a probabilidade de um indivíduo permanecer sem o evento até um dado tempo, permitindo representar graficamente o perfil de sobrevida da população estudada.

2.5. MODELOS DE FRAÇÃO DE CURA

Segundo Maller e Zhou (1996) e Vahidpour (2016), para incorporar a presença de sobreviventes de longo prazo em um modelo de sobrevivência, há duas formas a serem consideradas: os modelos de mistura (*standard cure rate models*), e os modelos de não-mistura (*non-mixture cure rate models*). A primeira categoria, a mais amplamente utilizada, consiste em modificar a função de sobrevivência como uma combinação ponderada entre a fração suscetível ao evento e a fração curada, refletindo explicitamente a heterogeneidade da população (LAMBERT, 2007; DE FREITAS E RODRIGUES, 2013; OLIVEIRA, 2019).

Já a segunda categoria, que são menos convencionais, ajustam a função de sobrevivência diretamente por meio da transformação da função de distribuição acumulada da distribuição *baseline*, de modo a representar a fração de indivíduos livres do risco após certo tempo (MAZUCHELI, COELHO-BARROS E ACHCAR, 2013; MARTINEZ, 2013, BAGHESTANI, 2015, OLIVEIRA, 2019).

Neste estudo, em particular, será considerada a abordagem de mistura para a construção do modelo de fração de cura com baseline Weibull. Especificamente, seja T uma variável aleatória positiva associada ao tempo até a ocorrência do evento.

Segundo Maller e Zhou (1996), o modelo de fração de cura de mistura assume que a probabilidade de o tempo até o evento ser superior a um tempo especificado t é dada por:

$$S(t) = \rho + (1 - \rho)S_0(t) \quad (2.6)$$

em que $\rho \in (0, 1)$ é o parâmetro de mistura que representa, em particular, a proporção de *sobreviventes de longo prazo* (ou indivíduos *não-suscetíveis*) e $S_0(t)$ denota uma função de sobrevivência para o grupo suscetível da população. Observe-se que, quando $t \rightarrow \infty$, tem-se $S(t) \rightarrow \rho$, ou seja, a função de sobrevivência apresenta uma assíntota no valor da fração de cura ρ . Essa propriedade permite que o modelo não apenas descreva o risco ao longo do tempo para os indivíduos suscetíveis, mas também forneça uma estimativa direta da proporção de pacientes efetivamente curados. Dessa forma, a interpretação da curva de sobrevivência torna-se mais informativa, permitindo diferenciar os efeitos temporais do risco do evento da existência de um subgrupo de longo prazo que não está mais sujeito ao desfecho estudado.

3 MATERIAIS E MÉTODOS

3.1. BASE DE DADOS

Para o desenvolvimento do presente estudo, foi utilizada a base de dados NKI Breast Cancer, proveniente do Netherlands Cancer Institute, obtida de forma secundária a partir do estudo de Van de Vijver et al. (2002). A base é composta por informações de 272 mulheres diagnosticadas com câncer de mama primário invasivo, cujos tumores apresentavam diâmetro inferior a 5 cm (classificação pT1 ou pT2), com diagnósticos realizados entre os anos de 1984 e 1995. As pacientes foram submetidas a mastectomia radical modificada ou cirurgia conservadora com dissecação dos linfonodos axilares, sendo a radioterapia administrada conforme indicação clínica. A base de dados inclui tantos pacientes com linfonodos negativos (pN0) quanto positivos (pN+). As principais características clínicas da amostra estão resumidas na Tabela 1.

Tabela 1. Características das 272 pacientes com câncer de mama do Instituto do Câncer da Holanda do estudo de Van de Vijver et al. (2002).

Variáveis Contínuas			Variáveis Categóricas		
Variável	Não-Censurados	Censurados	Variável	Não-Censurados	Censurados
	Média ± DP	Média ± DP		n (%)	n (%)
Diâmetro	25.38 ± 8.63	21.47 ± 8.47	Quimioterapia		
Linfonodos Positivos	1.59 ± 2.46	1.22 ± 1.94	Não	51 (67%)	111 (58%)
Idade	42.84 ± 6.2	44.55 ± 5.09	Sim	25 (33%)	81 (42%)
			Terapia Hormonal		
			Não	70 (92%)	162 (84%)
			Sim	6 (8%)	30 (16%)
			Tipo de Cirurgia		
			Conservativa	39 (51%)	110 (57%)
			Amputação	37 (49%)	82 (43%)
			Tipo Histológico		
			Tipo 1	72 (95%)	182 (95%)
			Tipo 2	4 (5%)	10 (5%)
			Grade do Tumor		
			Grau 1	4 (5%)	64 (33%)
			Grau 2	24 (32%)	71 (37%)
			Grau 3	48 (63%)	57 (30%)

Fonte: Elaboração própria com base nos dados.

3.2. DISTRIBUIÇÃO WEIBULL

De acordo com a literatura, a distribuição Weibull foi proposta em 1951 por Waloddi Weibull. Essa distribuição, em particular, é conhecida por seu vasto campo de aplicações em praticamente todas as áreas da ciência e sua capacidade de lidar com dados assimétricos, e sua flexibilidade em modelar diferentes formas de risco ao longo do tempo. Dependendo dos valores dos seus parâmetros, a distribuição pode modelar taxas de risco constantes, decrescentes ou crescentes, tornando-a muito adequada para diversas aplicações. Sua função de sobrevivência é descrita por:

$$S_0(t; \lambda, \gamma) = e^{-\lambda t^\gamma} \quad (3.1)$$

em que $\lambda > 0$ é o parâmetro de escala e $\gamma > 0$ é o parâmetro de forma. Para a construção do modelo de interesse, será considerada a função de sobrevivência acima como *baseline*. A escolha da distribuição Weibull, neste caso, se justifica pela natureza dos dados observacionais e pela presença de indícios empíricos de heterogeneidade no risco de morte entre os pacientes. A incorporação da fração de cura nesse modelo, em particular, possibilita capturar nuances clínicas importantes, como a identificação de um subgrupo de pacientes com insuficiência cardíaca que pode ter alcançado um estado de recuperação sustentada.

3.2.1. INCORPORAÇÃO DA FRAÇÃO DE CURA

Assumindo a estrutura do modelo de fração de cura de mistura, expressa em (1), pode-se incorporar a fração de cura à distribuição Weibull redefinindo sua função de sobrevivência como:

$$S(t; \lambda, \gamma, \rho) = \rho + (1 - \rho)e^{-\lambda t^\gamma} \quad (3.2)$$

em que $\rho \in (0, 1)$ é o parâmetro de mistura que representa, em particular, a proporção de *sobreviventes de longo prazo* (ou indivíduos *não-suscetíveis*). Para este modelo, o logaritmo da função de verossimilhança para $\theta = (\lambda, \gamma, \rho)$, com base em (2), é descrita, no contexto de análise de sobrevivência, por:

$$\begin{aligned} \ell(\lambda, \gamma, \rho) = & r \ln(1 - \rho) + r \ln \gamma + r \ln \lambda + \\ & + (\gamma - 1)v - \lambda A_1(\gamma) + A_2(\rho, \lambda, \gamma) \end{aligned} \quad (3.3)$$

tal que,

$$\begin{aligned} r &= \sum_{i=1}^n \delta_i \\ v &= \sum_{i=1}^n \delta_i \ln t_i \end{aligned} \quad (3.4)$$

$$A_1(\gamma) = \sum_{i=1}^n \delta_i t_i^\gamma$$

$$A_2(\rho, \lambda, \gamma) = \sum_{i=1}^n (1 - \delta_i) \ln(\rho + (1 - \rho)e^{-\lambda t_i^\gamma})$$

em que δ_i é o indicador de censura para o paciente i , e t_i e o tempo observado para o paciente i . Com base na verossimilhança acima, para determinar o estimador do vetor de parâmetros $\theta = (\lambda, \gamma, \rho)$, deve-se maximizar $\ell(\theta) = \log L(\theta)$. No entanto, dada a complexidade da verossimilhança, é necessário fazer o uso de métodos numéricos como Newton-Raphson ou Nelder-Mead para determinar os estimadores de máxima verossimilhança de θ . No presente trabalho, esse processo será realizado por meio da função *optim* do ambiente R (versão 4.3.0), usando o método numérico de Nelder-Mead para a maximização de $\ell(\theta)$. Sob condições de regularidade, a distribuição assintótica do estimador de máxima verossimilhança $\hat{\theta}$ é uma distribuição normal multivariada com média θ e matriz de covariância $\Sigma(\hat{\theta})$, a qual pode ser estimada de forma consistente pelo inverso da matriz de informação de Fisher, avaliada em $\theta = \hat{\theta}$, dada por:

$$\Sigma(\hat{\theta}) = \left[-\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^\top} \right]^{-1} \quad (3.5)$$

em que:

$$\begin{aligned} \frac{\partial \ell(\theta)}{\partial \lambda} &= \frac{r}{\lambda} - A_1(\gamma) + \frac{\partial A_2(\rho, \lambda, \gamma)}{\partial \lambda} \\ \frac{\partial \ell(\theta)}{\partial \gamma} &= \frac{r}{\gamma} + v - \lambda \frac{\partial A_1(\gamma)}{\partial \gamma} + \frac{\partial A_2(\rho, \lambda, \gamma)}{\partial \gamma} \\ \frac{\partial \ell(\theta)}{\partial \rho} &= -\frac{r}{1 - \rho} + \frac{\partial A_2(\rho, \lambda, \gamma)}{\partial \rho} \end{aligned} \quad (3.6)$$

3.2.2. INCORPORAÇÃO DE COVARIÁVEIS

Neste trabalho, o modelo de fração de cura proposto em (3.2), foi expandido para o contexto de regressão, incorporando um vetor de covariáveis $\mathbf{x} = (x_1, \dots, x_k)'$ no parâmetro de fração de cura, ρ , a partir da seguinte estrutura de regressão:

$$\ln\left(\frac{\rho_i}{1 - \rho_i}\right) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} \quad (3.6)$$

em que $\beta_i, i = 0, \dots, k$ são os parâmetros de regressão a serem estimados. As variáveis incluídas na análise foram: X_1 (Idade), X_2 (Quimioterapia), X_3 (Terapia Hormonal), X_4 (Tipo de Cirurgia), X_5 (Histologia do Tumor), X_6 (Diâmetro do Tumor), X_7 (Linfonodos Positivos), e X_8 (Grade do Tumor).

3.3. RESÍDUOS DE COX-SNELL

Após a estimação dos parâmetros em modelos de sobrevivência, a avaliação da adequação do ajuste (*goodness-of-fit*) é um passo fundamental. Os resíduos de Cox-Snell são uma ferramenta diagnóstica útil para este fim, pois permitem verificar se o modelo proposto captura de forma satisfatória a estrutura de risco dos dados. A construção desses resíduos baseia-se na teoria de que, se o modelo for corretamente especificado, o tempo de sobrevida de cada indivíduo, quando transformado pela função de risco cumulativo estimada do modelo, deverá seguir uma distribuição exponencial padrão com taxa de risco igual a 1 (COX & SNELL, 1968; KALBFLEISCH & PRENTICE, 2002). Em outras palavras, o resíduo de Cox-Snell para o i -ésimo indivíduo, r_{C_i} , é definido pela função de risco cumulativo (hazard cumulativo) estimada do modelo $\hat{H}(t_i)$:

$$r_{C_i} = \hat{H}(t_i) = H_0(t_i) \exp(x_i^T \hat{\boldsymbol{\beta}}) \quad (27)$$

em que t_i é o tempo de sobrevida observado, x_i é o vetor de covariáveis e $\hat{\boldsymbol{\beta}}$ são os coeficientes estimados. A avaliação do ajuste é realizada inspecionando a distribuição desses resíduos. Para um modelo bem ajustado, os r_{C_i} devem se comportar como uma amostra aleatória de uma distribuição exponencial. Isso é verificado por meio de

um gráfico de probabilidade cumulativa (*cumulative hazard plot*). Neste gráfico, a função de risco cumulativo dos resíduos, estimada por meio de um método não-paramétrico (como a estimativa de Nelson-Aalen), é plotada em função dos próprios resíduos. Se o modelo estiver bem ajustado, a curva resultante deverá se aproximar de uma linha reta com inclinação de 45 graus (eixo y igual ao eixo x), passando pela origem. Desvios significativos dessa linha reta indicam que o modelo não está adequadamente ajustado, sugerindo que a forma funcional da função de risco base ou a relação com as covariáveis pode estar incorreta.

Quando se considera a presença de uma fração de cura no modelo de sobrevivência, a definição e a interpretação dos resíduos de Cox-Snell precisam ser modificadas, pois parte da população é considerada imune ao evento de interesse, e, portanto, o pressuposto de que todos os indivíduos falham em algum instante não é válido. Nesse cenário, o tempo de sobrevivência observado t_i é modelado condicionalmente à probabilidade de suscetibilidade, e o resíduo de Cox-Snell deve incorporar a estrutura do modelo de cura. Para um indivíduo i , o resíduo passa a ser definido em termos da função de risco cumulativo condicional para os suscetíveis, ajustada pela probabilidade de cura, isto é,

$$r_{C_i} = -\log\left(\frac{\hat{S}(t_i)}{\hat{p}_i + (1 - \hat{p}_i)\hat{S}_0(t)^{\exp(x_i^T \hat{\beta})}}\right) \quad (28)$$

onde $\hat{S}(t_i)$ é a função de sobrevivência estimada, $\hat{p}_i$ representa a probabilidade estimada de cura para o indivíduo i , $\hat{S}_0(t)$ é a função de sobrevivência *baseline*, e $\hat{\beta}$ são os coeficientes estimados do modelo. A interpretação gráfica, neste caso, também sofre alteração. Como a presença da fração de cura implica que a função de sobrevivência não tende a zero, mas sim a um platô correspondente à proporção de curados, a função de risco cumulativo dos resíduos não seguirá uma distribuição exponencial padrão para toda a amostra. A avaliação do ajuste deve, portanto, ser restrita ao subconjunto de indivíduos suscetíveis. Para estes, a análise gráfica segue de maneira análoga ao caso clássico: a curva estimada do risco cumulativo dos resíduos deve se aproximar da reta com inclinação de 45 graus.

4 ESTUDO DE SIMULAÇÃO MONTE CARLO

Para avaliar as propriedades dos estimadores de máxima verossimilhança do modelo proposto, foi conduzido um estudo de simulação Monte Carlo. Para tal, foi considerada a geração de 1000 amostras pseudoaleatórias com tamanhos 25, 50, 75 e 100, a partir do modelo Weibull com fração de cura e aproximadamente 35% de censura à direita. O método da transformação inversa foi utilizado para gerar as amostras pseudoaleatórias, e os valores nominais considerados para os parâmetros do modelo foram organizados em três 3 cenários. No primeiro cenário consideramos $\lambda = 2.0$, $\gamma = 1.5$ e $\rho = 0.3$, representando um risco crescente moderado.

O segundo cenário, definido por $\lambda = 1.5$, $\gamma = 1.0$ e $\rho = 0.1$, representa um processo com risco constante e baixa proporção de cura.

Já o terceiro cenário, com $\lambda = 2.5$, $\gamma = 2.5$ e $\rho = 0.5$, caracteriza um risco crescente mais acentuado e alta proporção de cura.

Esses cenários permitem avaliar o comportamento dos estimadores em diferentes condições de risco e censura.

Como estimador pontual, adotou-se o estimador de máxima verossimilhança, cujo desempenho foi avaliado em termos de viés (B) e erro quadrático médio (EQM). Assim, a partir das amostras dos parâmetros do modelo, obtiveram-se estimativas de Monte Carlo aproximadas para essas medidas, dadas por:

$$\hat{B}(\hat{\epsilon}) = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{\epsilon}_i - \epsilon)$$

E,

$$\widehat{EQM}(\hat{\epsilon}) = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{\epsilon}_i - \epsilon)^2$$

onde $\epsilon = (\lambda, \gamma, \rho)$ representa os parâmetros do modelo proposto e $\{\hat{\epsilon}_i\}_{i=1}^{1000}$ é a sequência das estimativas obtidas para cada amostra. Os resultados são apresentados na Tabela 2, 3 e 4 e mostram que o desempenho dos estimadores melhora sistematicamente com o aumento do tamanho amostral, independentemente dos valores adotados para os parâmetros. Em todos os cenários, tanto o viés quanto o EQM diminuem conforme n cresce, indicando que os estimadores são consistentes.

Tabela 2. Resultado do estudo de simulação Monte Carlo: $\lambda = 2.0$, $\gamma = 1.5$ e $\rho = 0.3$.

n	Parâmetro	B (Viés)	EQM
25	λ	0.0147	0.1339
	γ	0.1689	0.1787
	ρ	-0.0022	0.0106
50	λ	0.0041	0.0755
	γ	0.054	0.0544
	ρ	-0.0023	0.0049
75	λ	-0.0012	0.0456
	γ	0.0471	0.0393
	ρ	0.0010	0.0033
100	λ	0.0024	0.0352
	γ	0.0341	0.0264
	ρ	-0.0026	0.0026

Tabela 3. Resultado do estudo de simulação Monte Carlo: $\lambda = 1.5$, $\gamma = 1.0$ e $\rho = 0.1$.

n	Parâmetro	B (Viés)	EQM
25	λ	0.0022	0.1390
	γ	0.0974	0.0618
	ρ	0.0140	0.0038
50	λ	0.0134	0.0749
	γ	0.0286	0.0197
	ρ	0.0019	0.0022
75	λ	0.0074	0.0478
	γ	0.0226	0.0137
	ρ	0.0000	0.0016
100	λ	0.0124	0.0353
	γ	0.0206	0.0102
	ρ	0.0000	0.0012

Tabela 4. Resultado do estudo de simulação Monte Carlo: $\lambda = 2.5$, $\gamma = 2.5$ e $\rho = 0.5$.

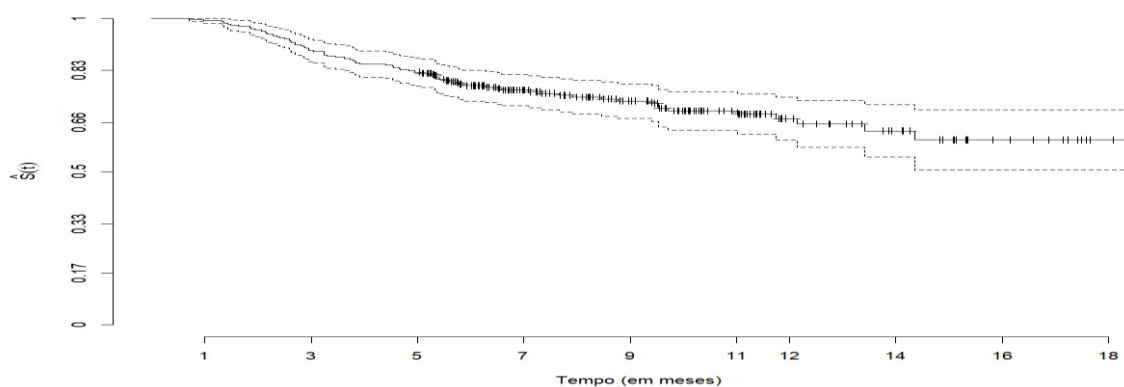
n	Parâmetro	B (Viés)	EQM
25	λ	-0.0314	0.1231
	γ	0.4346	0.9631
	ρ	-0.0075	0.0125
50	λ	-0.0174	0.0554
	γ	0.1938	0.3196
	ρ	-0.0011	0.0068
75	λ	-0.0008	0.0384
	γ	0.1110	0.1644
	ρ	0.0010	0.0043
100	λ	-0.0042	0.0274
	γ	0.0833	0.1150
	ρ	-0.0009	0.0029

5. RESULTADOS E DISCUSSÕES

5.1. ANÁLISE EXPLORATORIA DE DADOS

Como etapa inicial da análise, foi construída a curva de sobrevivência, ao longo do tempo de acompanhamento do estudo, estimada pelo estimador empírico de Kaplan-Meier, sendo apresentada na Figura 2. Nessa figura, a linha sólida representa a estimativa pontual da função de sobrevivência, enquanto a faixa sombreada corresponde ao intervalo de confiança de 95%, fornecendo uma medida da incerteza associada às estimativas em cada ponto temporal.

Figura 2. Curva de sobrevivência estimada via Kaplan-Meier referente ao tempo de sobrevida dos pacientes com câncer de mama.

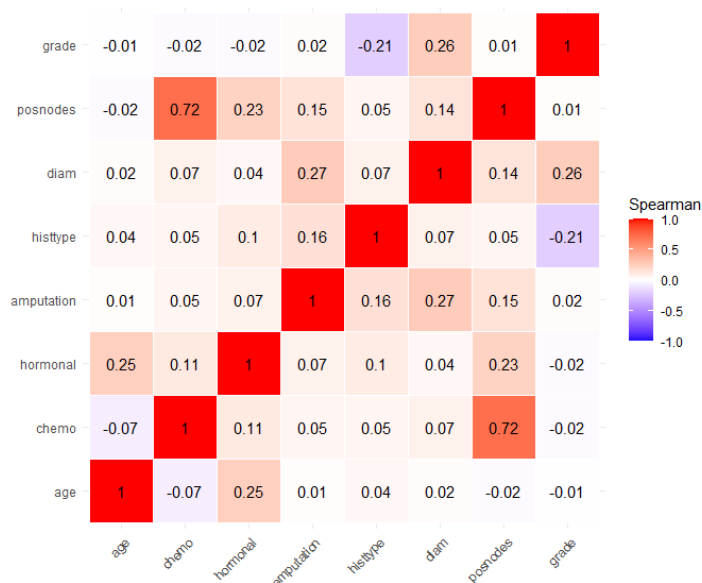


Fonte: Elaboração própria com base nos dados.

Os resultados evidenciam uma queda acentuada na probabilidade de sobrevivência nos primeiros anos após o início do acompanhamento, seguida de uma estabilização da curva, formando um platô em aproximadamente 0,604, indicando que 60,4% dos pacientes permaneceram livres do evento ao longo do período de acompanhamento (isto é, há presença de remissão a longo prazo). Esse resultado sugere que modelos de sobrevivência tradicionais, como o modelo de riscos proporcionais de Cox — os quais assumem que todos os indivíduos permanecem indefinidamente em risco — podem ser inadequados para a análise, o que motiva a adoção de um modelo com fração de cura, como o modelo de fração de cura proposto.

A Figura 2 exibe a matriz de correlação entre as covariáveis consideradas no estudo, obtida a partir do coeficiente de correlação de Spearman, o qual é apropriado para avaliar associações monotônicas, especialmente com a presença de variáveis categóricas. Os coeficientes variam entre -1 e 1, sendo representados por um gradiente de cores, onde tons azulados indicam correlações negativas e tons avermelhados, correlações positivas. Da Figura 2, observa-se uma forte correlação positiva entre a variável *chemo* (indicação de quimioterapia) e *posnodes* (presença de linfonodos positivos), com $\rho_{\text{Spearman}} = 0,72$. Este resultado é condizente com a prática clínica, dado que a presença de linfonodos positivos está fortemente associada à indicação de quimioterapia adjuvante. As demais correlações são, em geral, fracas ou desprezíveis, com coeficientes próximos de zero, indicando ausência de associação monotônica relevante entre essas variáveis.

Figura 3. Matriz de correlação de Spearman entre as covariáveis consideradas no estudo.

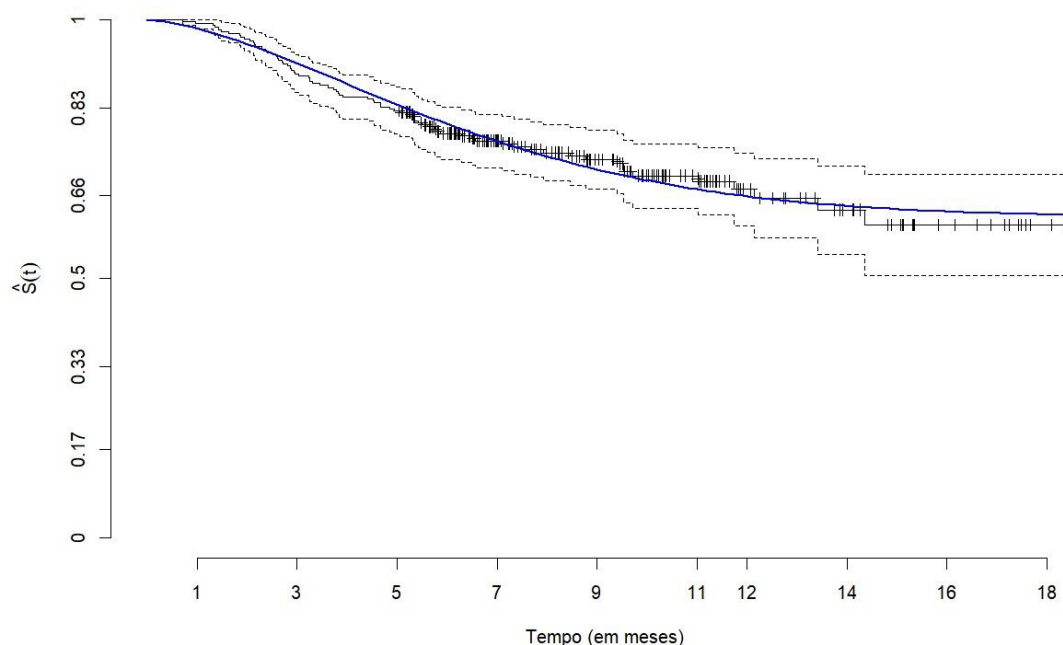


Fonte: Elaboração própria com base nos dados.

5.2. ANÁLISE PELO MODELO DE FRAÇÃO DE CURA WEIBULL

Na análise inicial, foi realizada a estimação dos parâmetros do modelo proposto sem considerar o uso de estruturas de regressão, com o intuito de avaliar se o modelo proposto se adequava a curva de sobrevivência empírica obtida pelo estimador Kaplan-Meier. O resultado dessa análise mostrou que o processo de estimação dos parâmetros do modelo proposto convergiu de forma satisfatória, e a fração de cura estimada foi de 61,8%. A Figura 3 ilustra a comparação gráfica entre o modelo de fração de cura de mistura Weibull proposto e a curva de sobrevivência estimada via Kaplan-Meier, onde observa-se que a curva paramétrica do modelo proposto aderem bem à trajetória empírica dos dados, sugerindo que a distribuição Weibull (com fração de cura) captura adequadamente a dinâmica de sobrevivência da coorte.

Figura 4. Curva de sobrevivência empírica (estimada via Kaplan-Meier) vs curva de sobrevivência ajustada pelo modelo de fração de cura de mistura Weibull para o tempo de sobrevivência dos pacientes com câncer de mama.



Fonte: Elaboração própria com base os dados.

Dando continuidade ao estudo, incorporou-se uma estrutura de regressão para o parâmetro ρ , o qual representa a fração de cura do modelo proposto, com o objetivo de avaliar as influências dos fatores de risco sobre o comportamento dos pacientes curados. Os fatores de risco considerados na análise foram: a idade do paciente, realização de quimioterapia, realização de tratamento hormonal, tipo de cirurgia realizada, tipo histológico do tumor, diâmetro do tumor, número de linfonodos positivos, e a grade do tumor - que foi reclassificada em duas categorias: risco moderado (junção da grade 1 e grade 2) e risco crítico (grade 3). Os resultados da análise são expostos na Tabela 2, em que são apresentadas as estimativas dos parâmetros, os erros-padrão associados a essas estimativas, e os intervalos de confiança a 95% para os efeitos das variáveis explicativas sobre a probabilidade de cura de pacientes com câncer de mama, utilizando o modelo proposto de fração de cura de mistura Weibull.

Tabela 5. Estimativas de parâmetros e intervalos de confiança para 95% do modelo de regressão de fração de cura proposto.

Parâmetro	Estimativa	Erro-Padrão	I.C de 95%
γ	1.6071	0.1668	(1.2803, 1.9340)
λ	6.6143	0.7395	(5.1649, 8.0636)
β_0 (Intercepto)	2.4900	1.8486	(-1.1331, 6.1131)
β_1 (Idade)	0.0460	0.0313	(-0.0153, 0.1073)
β_2 (Quimioterapia – Sim)	0.7705	0.4285	(-0.0692, 1.6103)
β_3 (Terapia Hormonal – Sim)	0.6212	0.6052	(-0.5650, 1.8074)
β_4 (Tipo de Cirurgia - Amputação)	-0.1057	0.3534	(-0.7983, 0.5869)
β_5 (Histologia do Tumor – Tipo 2)	-1.1879	0.8978	(-2.9475, 0.5718)
β_6 (Diâmetro do Tumor)	-0.0332	0.0210	(-0.0745, 0.0080)
β_7 (Linfonodos Positivos)	-0.1557	0.1050	(-0.3615, 0.0502)
β_8 (Grade do Tumor – Risco Crítico)	-2.3639	0.6246	(-3.5882, -1.1396)

Fonte: Elaboração própria com base os dados.

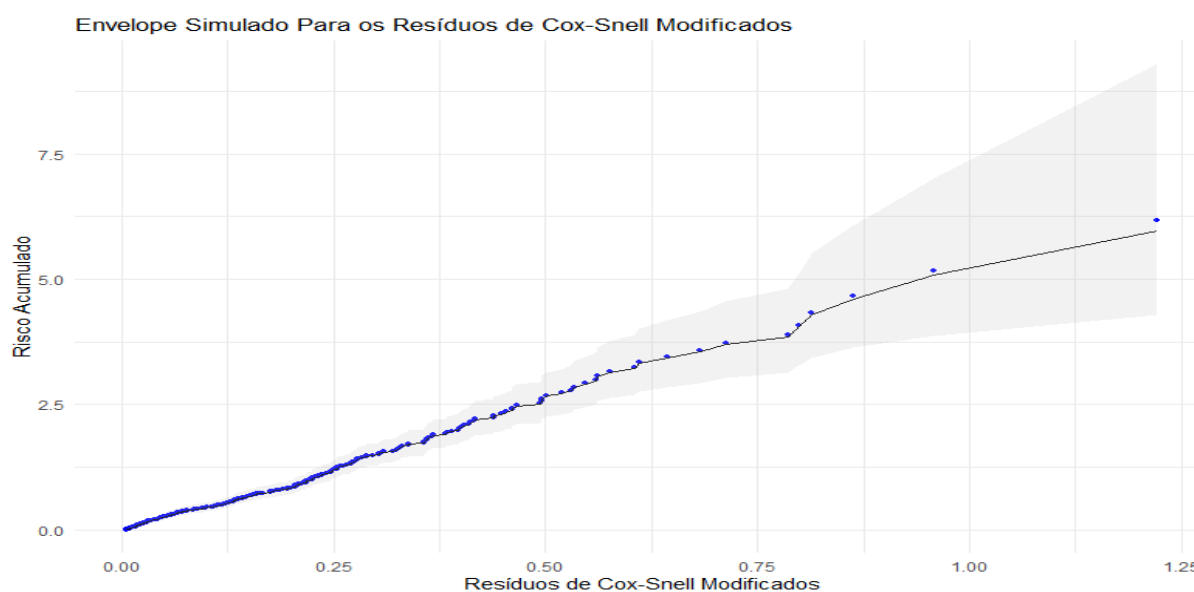
Dos resultados na Tabela 2, observa-se que apenas a variável correspondente a grade do tumor foi significativa para o estudo (intervalo de confiança não contém o zero). Esse resultado indica que pacientes com tumores classificados como de grau de *risco crítico* têm menor chance de pertencer à fração de cura, isto é, estão mais propensos a integrar o grupo suscetível ao evento de falha (óbito). Esse efeito, em particular, implica em uma redução na chance de remissão completa do câncer.

Em relação às demais variáveis incluídas no modelo, observa-se que, embora não tenham apresentado significância estatística, obtêm efeitos clínicos. A variável idade apresentou coeficiente positivo, sugerindo que o aumento da idade tende a reduzir a probabilidade de pertencer à fração de cura, resultado que está em conformidade com estudos que apontam a idade avançada como um fator associado a menor resposta terapêutica e maior risco de recorrência (Ferlay et al., 2024; Lambert, 2007). Para as variáveis relacionadas ao tratamento, como quimioterapia e terapia hormonal, ambos os efeitos estimados foram positivos, indicando uma possível associação ao evento de falha. Essa relação pode ser explicada pelo fato de que tais terapias são geralmente indicadas a pacientes com tumores em estágios mais avançados ou de maior agressividade, o que naturalmente reduz a probabilidade de

cura, conforme discutido por Shabankhani et al. (2022) e Van de Vijver et al. (2002). O tipo de cirurgia também não se mostrou significativo, mas apresentou coeficiente negativo, o que sugere uma leve tendência de melhora na chance de cura em pacientes submetidas à amputação. Já as variáveis histologia do tumor, diâmetro do tumor e número de linfonodos positivos, embora também não significativas, revelaram comportamentos coerentes com a literatura: histologias menos agressivas e tumores de menor tamanho indicam melhor prognóstico, enquanto a presença de linfonodos positivos tende a reduzir a probabilidade de cura (Vahidpour et al., 2016; Kalbfleisch & Prentice, 2002). Assim, mesmo sem significância estatística, os sinais e magnitudes das estimativas corroboram os achados prévios sobre fatores prognósticos no câncer de mama, reforçando a consistência e consistência biológica do modelo ajustado.

Por fim, na Figura 5, é apresentado o envelope simulado para os resíduos de Cox-Snell para o modelo proposto. Observa-se que os pontos seguem de forma razoável a linha de referência, permanecendo majoritariamente dentro da região sombreada do envelope simulado. Isso indica uma boa aderência do modelo aos dados, sugerindo que não há evidências fortes de violação dos pressupostos do modelo de regressão ajustado.

Figura 5. Resíduos de Cox-Snell para o modelo de fração de cura de mistura Weibull proposto.



Fonte: Elaboração própria com base os dados.

6. CONCLUSÃO

O presente estudo teve como objetivo principal analisar os fatores associados à sobrevivência de pacientes com câncer de mama, com ênfase na identificação de variáveis clínicas relevantes, por meio da aplicação de um modelo de fração de cura do tipo mistura baseado na distribuição Weibull. Inicialmente, a análise sem estrutura de regressão evidenciou a adequação do modelo proposto para representar a dinâmica de sobrevivência observada, como indicado pela boa concordância entre a curva estimada pelo modelo paramétrico e a curva empírica de Kaplan–Meier. A fração de cura estimada foi de 61,8%, sugerindo que uma parcela expressiva das pacientes atingiu remissão completa, permanecendo livre do evento durante o período de acompanhamento.

Ao incorporar uma estrutura de regressão para modelar a fração de cura em função de variáveis clínicas, foi possível avaliar o impacto de potenciais fatores prognósticos sobre a probabilidade de remissão de longo prazo. Dentre os preditores analisados, apenas a variável relacionada ao grau do tumor (grau crítico) apresentou associação estatisticamente significativa, indicando que pacientes com tumores de maior agressividade apresentam menor chance de cura. As demais covariáveis não mostraram significância estatística, o que pode estar relacionado a limitações amostrais, como o tamanho da amostra ou à colinearidade entre variáveis clínicas. Adicionalmente, a análise baseada nos resíduos de Cox–Snell adaptados ao modelo de mistura confirmou a adequação do ajuste, não sendo observadas violações relevantes dos pressupostos.

REFERÊNCIAS

- ABEBE, T. B. et al.** The prognosis of heart failure patients: Does sodium level play a significant role? *PLoS ONE*, v. 13, n. 11, p. e0207242, 2018.
- ACHCAR, J. A.; BOLFARINE, H.; RODRIGUES, J.** Cure rate models and their applications in survival analysis. *Lifetime Data Analysis*, v. 18, n. 4, p. 457–475, 2012. DOI: <https://doi.org/10.1007/s10985-012-9224-7>.
- ACHCAR, J. A. et al.** Cure fraction models using mixture and non-mixture distributions: a Bayesian approach. *Computational Statistics & Data Analysis*, v. 56, n. 5, p. 1377–1392, 2012.
- AMERICAN CANCER SOCIETY.** Breast cancer in men. 2023. Disponível em: <https://www.cancer.org/cancer/breast-cancer-in-men.html>. Acesso em: 4 out. 2024.
- BAGHESTANI, A. R.; ABDOLLAHZADEH, S.; MAZLOOMI, S.** A cure model in survival analysis for breast cancer patients in presence of competing risks. *Asian Pacific Journal of Cancer Prevention*, v. 16, n. 18, p. 8399–8404, 2015.
- BERNARDO, W.; SIMÕES, R.; SILVINATO, A.** BRCA1 e BRCA2 em câncer de mama. [s.l.]: [s.n.], 2016. Disponível em: <https://amb.org.br/wp-content/uploads/2021/08/BRCA1-E-BRCA-2-FINAL-2016.pdf>. Acesso em: 18 out. 2024.
- BORGES, P.** Novos modelos de sobrevivência com fração de cura baseados no processo de carcinogênese. 2012. Dissertação (Mestrado em Estatística) – Universidade Federal de São Carlos, São Carlos, 2012. Disponível em: <https://repositorio.ufscar.br/bitstream/handle/ufscar/4489/4552.pdf?sequence=1&isAllowed=y>. Acesso em: 26 nov. 2024.
- BRAY, F. et al.** Global cancer transitions according to the Human Development Index (2008–2030): A population-based study. *The Lancet Oncology*, v. 19, n. 7, p. 790–801, 2012. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/22658655/>. Acesso em: 5 out. 2024.
- CARDOSO, F. et al.** 4th ESO-ESMO International Consensus Guidelines for Advanced Breast Cancer (ABC 4). *Annals of Oncology*, n. 12, p. 1634-1657, 2018. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/30032243/>. Acesso em: 04 out. 2024.
- CARVALHO, Marília Sá. et al.** Análise de sobrevivência: teoria e aplicações em saúde. 2. ed., 2019. 432 p.
- CLAUDINO, J. E.; OLIVEIRA, R. H. de.** Modelo de fração de cura na análise de sobrevida de pacientes com câncer de pele. Curitiba, 2013. Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Federal do Paraná, Curitiba, 2013. Disponível em: https://coordest.ufpr.br/wp-content/uploads/2018/12/TCC_JoseaneRafaela.pdf. Acesso em: 26 nov. 2024.

COLLABORATIVE GROUP ON HORMONAL FACTORS IN BREAST CANCER.

Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118,964 women with breast cancer from 117 epidemiological studies. *Lancet Oncology*, v. 13, p. 1141–1151, 2012.

COLLETT, D. Modelling Survival Data in Medical Research. 3. ed. Chapman & Hall/CRC, 2015.

COX, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 34, n. 2, p. 187–220, 1972.

COX, D. R.; SNELL, E. J. A General Definition of Residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 30, p. 248–275, 1968.

CRESTANI, F.; Jorge, L. Aplicações de modelo de fração de cura em pacientes durante o tratamento . 2023. Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Estadual Paulista, São Paulo, 2023. Disponível em: <https://repositorio.unesp.br/items/e6b8ef4d-57c9-4f00-80c1-48eef00b7350> Acesso em: 26 nov. 2024.

DELEN, D.; WALKER, G.; KADAM, A. Prevendo a sobrevivência do câncer de mama: uma comparação de três métodos de mineração de dados. *Artificial Intelligence in Medicine*, v. 34, p. 113–127, 2005. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0933365704001010>. Acesso em: 22 set. 2024.

DE FREITAS, L. A.; RODRIGUES, J. Standard exponential cure rate model with informative censoring. *Communications in Statistics-Simulation and Computation*, v. 42, n. 1, p. 8-23, 2013.

DE LA CRUZ, R.; FUENTES, C.; PADILLA, O. A Bayesian mixture cure rate model for estimating short-term and long-term recidivism. *Entropy*, v. 25, n. 1, p. 56, 2023. DOI: 10.3390/e25010056. Disponível em: <https://www.mdpi.com/1099-4300/25/1/56>

FERLAY, J. et al. Global Cancer Observatory: Cancer Today (GLOBOCAN 2024). Lyon: International Agency for Research on Cancer (IARC), 2024. Disponível em: <https://gco.iarc.fr/today>.

FERRAZ, R.; MOREIRA-FILHO, D. C. Análise de sobrevivência de mulheres com câncer de mama: modelos de riscos competitivos. *Ciência & Saúde Coletiva* [online]. 2017, v. 22, n. 11, pp. 3743-3754. Disponível em: <https://doi.org/10.1590/1413-812320172211.05092016>. Acesso em: 22 set. 2024.

FOULKES, W. D. Inherited susceptibility to common cancers. *New England Journal of Medicine*, v. 359, n. 20, p. 2143-2153, 2008. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/19005198/> . Acesso em: 04 out. 2024

GALLARDO, D. I.; BRANDÃO, M.; LEÃO, J.; BOURGUIGNON, M.; CALSAVARA, V. Um novo modelo de mistura com taxa de cura aplicado a dados de câncer de

mama. *Biometrical Journal*, v. 66, n. 6, set. 2024. DOI: 10.1002/bimj.202300257. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/39104134>

GIOLO, S. R.; COLOSIMO, E. A. *Análise de Sobrevida Aplicada*. São Paulo: Edgard Blücher, 2006.

GORDON, N. H. *Statistical Methods for the Analysis of Survival Data*. New York: Chapman & Hall, 1996.

HASSANPOUR, H.; DEGHANI, M. Cancer: A Genetic Disease. *Medical Journal of the Islamic Republic of Iran*, v. 31, p. 122, 2017. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5651800/>. Acesso em: 18 out. 2024.

IARC (International Agency for Research on Cancer). GLOBOCAN 2020: New Global Cancer Data. 2020. Disponível em: <https://gco.iarc.fr/today/home>. Acesso em: 18 out. 2024.

INSTITUTO NACIONAL DE CÂNCER (INCA). Câncer de mama. Instituto Nacional de Câncer. 2022. Disponível em: <https://www.inca.gov.br/tipos-de-cancer/cancer-de-mama>. Acesso em: 4 out. 2024.

KALBFLEISCH, J. D.; PRENTICE, R. L. *The Statistical Analysis of Failure Time Data*. 2. ed. New York: Wiley, 2002.

KLEIN, J. P.; MOESCHBERGER, M. L. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 1997.

KLEINBAUM, D. G.; KLEIN, M. *Survival Analysis: A Self-Learning Text*. Springer, 2012.

LAMBERT, P. C. Modeling of the cure fraction in survival studies. *The Stata Journal*, v. 7, n. 3, p. 351-375, 2007.

LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. Wiley-Interciência, 1982.

LI, J.; MA, S. *Survival Analysis in Medicine and Genetics*. 1. ed. New York: CRC Press, 2013. Disponível em: <https://www.elsevier.es/en-revista-allergologia-et-immunopathologia-105linkresolver-survival-analysis-cox-regression-S0301054611002679>. Acesso em: 22 set. 2024.

LIMA, I. da S.; SILVA, S. J. da; BRIGHENTI, C. R. G.; NAKAMURA, L. R.; OLIVEIRA, T. A.; OLIVEIRA, M. E. C.; RAMIRES, T. G. Análise do tempo de sobrevida em mulheres com câncer de mama por modelos de regressão distribucional. *Cadernos de Saúde Pública*, v. 41, n. 8, e00073324, 2025. DOI: 10.1590/0102-311XEN073324. Disponível em: <https://www.scielo.org/article/csp/2025.v41n8/e00073324>

MALLER, R. A.; ZHOU, X. *Survival Analysis with Long-Term Survivors*. Wiley, 1997.

MARTINEZ, E. Z. et al. Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data. *Computer methods and programs in biomedicine*, v. 112, n. 3, p. 343-355, 2013.

MAZUCHELI, J.; COELHO-BARROS, E. A.; ACHCAR, J. A. The exponentiated exponential mixture and non-mixture cure rate model in the presence of covariates. *Computer methods and programs in biomedicine*, v. 112, n. 1, p. 114-124, 2013.

MOORE, Dirk F. Applied survival analysis using R. 1ª ed. New York: Springer, 2016.

MITCHELL, T. M. Machine Learning. 1. ed. New York: McGraw-Hill Science/Engineering/Math, 1997. 432 p.

OLIVEIRA, R. P.; MENEZES, A. F. B.; MAZUCHELI, J.; ACHCAR, J. A. Mixture and nonmixture cure fraction models assuming discrete lifetimes: Application to a pelvic sarcoma dataset. *Biometrical Journal*, [S.l.], v. 61, n. 5, p. 1281–1297, 2019.

OLIVEIRA, R. P. Multivariate lifetime models to evaluate long-term survivors in medical studies. 2019. Tese de Doutorado. Universidade de São Paulo.

OLIVEIRA, R. P.; PERES, M. V. de O.; BERTOLI, W.; ACHCAR, J. A. A non-mixture cure rate model for analyzing the survival lifetimes of cardiovascular heart failure patients. *Model Assisted Statistics and Applications*, v. 18, n. 1, p. 45–55, 2023.

PRICE, D. L.; MANATUNGA, A. K. Modelling survival data with a cured fraction using nonparametric methods. *Statistics in Medicine*, v. 20, n. 4, p. 535–546, 2001.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized Additive Models for Location, Scale and Shape (GAMLSS). *Applied Statistics*, v. 54, p. 507–554, 2005.

RODRIGUES, A.; BORGES, P.; SANTOS, B. A defective cure rate quantile regression model for male breast cancer data. *Journal of Applied Statistics*, v. 52, n. 8, p. 1485-1512, 14 nov. 2024. DOI: 10.1080/02664763.2024.2428272. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/02664763.2024.2428272>

SANTOS, T. M.; IZBICKI, R. Aprendizado de Máquina: uma abordagem estatística. 1. ed. São Paulo: UICLAP, 2022. 270 p.

SHABANKHANI, B.; EBADI, A.; MOGHIMBEIGI, A. Extended cure rate models: A comprehensive review. *Statistical Methods in Medical Research*, v. 31, n. 10, p. 1887–1911, 2022.

SILVA, B. L. S. Modelo de Regressão Kumaraswamy Log-Logístico para dados. 2019. Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade de Brasília, Brasília, 2019. Disponível em: https://bdm.unb.br/bitstream/10483/30122/1/2019_BeatrizLealSimoessilva_tcc.pdf . Acesso em: 30 out. 2024.

SOARES, J. F.; SIQUEIRA, A. L. Introdução à estatística médica. 1. ed. Belo Horizonte: Departamento de Estatística, UFMG, 1999. 300p.

SOUZA, H. C. C. *O Modelo Weibull Modificado Exponenciado de Longa Duração aplicado à sobrevida do câncer de mama*. 2015. 92 f. Dissertação (Mestrado em Saúde) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2015.

SUNG, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209-249, 2021. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/33538338/>. Acesso em: 05 out. 2024.

THERNEAU, T. M.; GRAMBSCH, P. M. Modeling Survival Data: Extending the Cox Model. Springer, 2000.

TUTZ, G.; SCHMID, M. Modeling Discrete Time-to-Event Data. New York: Springer, 2016.

UNGER-SALDAÑA, K. Challenges to the early diagnosis and treatment of breast cancer in developing countries. *World Journal of Clinical Oncology*, v. 5, n. 3, p. 465-477, 2014. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/25114860/>. Acesso em: 4 out. 2024.

VAHIDPOUR, N.; BAGHESTANI, A. R.; YAZDANI, J. Comparison of mixture and non-mixture cure models with application to gastric cancer data. *Asian Pacific Journal of Cancer Prevention*, v. 17, n. 7, p. 3319–3323, 2016.

VAN DE VIJVER, M. J. et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, v. 347, n. 25, p. 1999-2009, 2002. Disponível em: <https://www.nejm.org/doi/full/10.1056/NEJMoa021967>. Acesso em: 18 out. 2024.

WANG, J.; WU, S. Câncer de mama: uma visão geral das estratégias terapêuticas atuais, desafios e perspectivas. *Breast Cancer: Targets and Therapy*, v. 15, p. 721–730, 20 out. 2023. DOI: 10.2147/BCTT.S432526. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10596062/>. Acesso em: 17 out. 2024

WEINBERG, R. A. The biology of cancer. 2. ed. New York: Garland Science, 2013.

WIENKE, A.; LAMBERT, P. C.; SLOTTJE, P. A shared frailty model with a cure fraction. *Statistics in Medicine*, v. 25, p. 5375–5386, 2006.