



VANESSA MACULAN SILVÉRIO

**MODELO LOG-NORMAL COM FRAÇÃO DE CURA NO
CONTEXTO DA REGRESSÃO DISTRIBUCIONAL**

LAVRAS – MG

2025

VANESSA MACULAN SILVÉRIO

**MODELO LOG-NORMAL COM FRAÇÃO DE CURA NO CONTEXTO DA
REGRESSÃO DISTRIBUCIONAL**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Prof. Dr. Luiz Ricardo Nakamura
Orientador

Prof. Dr. Thiago Gentil Ramires
Coorientador

**LAVRAS – MG
2025**

**Ficha Catalográfica elaborada pelo Sistema de Geração
de Ficha Catalográfica da Biblioteca Universitária da UFLA, com
dados informados pelo(a) próprio(a) autor(a).**

Silvério, Vanessa Maculan.

Modelo log-normal com fração de cura no contexto da regressão distribucional /
Vanessa Maculan Silvério. - 2025.

73 p. : il.

Orientador: Luiz Ricardo Nakamura

Coorientador: Thiago Gentil Ramires

Dissertação (Mestrado Acadêmico) - Universidade Federal de Lavras, 2025.

Bibliografia.

1. Gamlss. 2. Análise de sobrevivência. 3. Censura. I. Nakamura, Luiz Ricardo
. II. Ramires, Thiago Gentil. III. Universidade Federal de Lavras. IV. Título.

VANESSA MACULAN SILVÉRIO

**MODELO LOG-NORMAL COM FRAÇÃO DE CURA NO CONTEXTO DA
REGRESSÃO DISTRIBUCIONAL**

**LOG-NORMAL CURE MODEL IN THE CONTEXT OF DISTRIBUTIONAL
REGRESSION**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 17 de junho de 2025.

Prof. Dr. Geraldo Magela da Cruz Pereira UFLA

Prof. Dr. Tiago Almeida de Oliveira UEPB

Prof. Dr. Luiz Ricardo Nakamura
Orientador

Prof. Dr. Thiago Gentil Ramires
Coorientador

**LAVRAS – MG
2025**

AGRADECIMENTOS

A Deus, por guiar meus passos, conceder-me força, saúde e sabedoria em todas as etapas desta caminhada.

Aos meus pais, Claudinei e Luciléia, por todo amor e por sempre acreditarem em mim, incentivando-me em todos os momentos da minha vida.

Ao meu esposo, Roger, pelo amor, apoio e compreensão, estando ao meu lado em todos os momentos.

Ao meu irmão, Julio César, pela amizade, apoio e incentivo sempre presentes.

Ao meu orientador, Luiz Ricardo Nakamura, e ao meu coorientador, Thiago Gentil Ramires, pela orientação, dedicação, confiança e ensinamentos essenciais para a realização deste trabalho.

Aos professores que participaram das bancas de qualificação e defesa, pelas contribuições importantes para o desenvolvimento desta pesquisa.

Ao Departamento de Estatística e à Universidade Federal de Lavras (UFLA), pela oportunidade, suporte e contribuição para minha formação acadêmica e pessoal.

Aos amigos e familiares que, de alguma maneira, contribuíram para a realização desta conquista.

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG).

A mente que se abre a uma nova ideia jamais retorna ao seu tamanho original.

(Albert Einstein)

RESUMO

O câncer é uma das principais causas de morbidade e mortalidade em todo o mundo, com grande impacto na saúde pública e na qualidade de vida dos pacientes diagnosticados. A análise de sobrevivência é uma ferramenta importante para entender o tempo até a ocorrência de eventos, como falecimento ou progressão da doença, sendo fundamental em estudos sobre câncer, especialmente quando envolvem dados censurados e fração de cura. Para construir modelos que representem esses dados, são comumente utilizados modelos probabilísticos, como as distribuições Weibull e log-normal. No entanto, modelos mais flexíveis, como os modelos aditivos generalizados para locação, escala e forma (GAMLSS), são capazes de oferecer uma modelagem mais precisa, especialmente quando os dados apresentam características complexas que não podem ser totalmente capturadas pelos modelos tradicionais. Assim, esta dissertação propõe a utilização do modelo log-normal com fração de cura dentro da estrutura dos GAMLSS. São apresentadas a formalização matemática do modelo, sua implementação no pacote `gamlss` do software R e sua aplicação em dados sobre câncer de pâncreas. Como resultado, a implementação da distribuição log-normal com fração de cura no contexto dos GAMLSS, aplicada a dados sobre câncer de pâncreas mostrou-se adequada, demonstrando que a metodologia adotada é eficaz para captar as particularidades da sobrevivência nessa doença. Além disso, o modelo permitiu a inclusão de covariáveis que explicam simultaneamente a mediana de sobrevivência, a variabilidade dos tempos de vida e a chance de cura, evidenciando seu potencial para aprimorar a análise de sobrevida em contextos clínicos e epidemiológicos.

Palavras-chave: GAMLSS; análise de sobrevivência; censura.

ABSTRACT

Cancer remains one of the leading causes of morbidity and mortality worldwide, exerting a substantial impact on public health and on the quality of life of affected patients. Survival analysis constitutes a fundamental tool for investigating the time until the occurrence of specific events, such as death or disease progression, and plays a pivotal role in cancer research, particularly when dealing with censored data and the presence of a cure fraction. Traditionally, probabilistic models such as the Weibull and log-normal distributions have been employed to represent such data. Nevertheless, more flexible approaches, such as generalized additive models for location, scale, and shape (GAMLSS), provide enhanced modeling capabilities, especially in scenarios where data exhibit complex features that are not adequately captured by conventional models. In this context, the present dissertation proposes the use of the log-normal cure model within the GAMLSS framework. The mathematical formalization of the model, its implementation in the `gamlss` package of the R software, and its application to pancreatic cancer data are presented. The results indicate that the implementation of the log-normal distribution with a cure fraction in the GAMLSS context is well-suited to the data, demonstrating the effectiveness of the proposed methodology in capturing the specific survival dynamics of this disease. Moreover, the model enables the incorporation of covariates that simultaneously explain median survival, variability in survival times, and the probability of cure, thereby underscoring its potential to advance survival analysis in both clinical and epidemiological research settings.

Keywords: GAMLSS; survival analysis; censoring.

INDICADORES DE IMPACTO

Esta dissertação apresenta um modelo estatístico flexível para a análise de sobrevivência em pacientes com câncer de pâncreas, uma das principais causas de morbidade e mortalidade no mundo. O estudo contribui para a saúde pública ao possibilitar estimativas mais precisas do tempo de sobrevivência, apoiar decisões clínicas, identificar fatores prognósticos relevantes e favorecer a melhoria da qualidade de vida dos pacientes. Do ponto de vista tecnológico, a implementação do modelo log-normal com fração de cura na estrutura dos GAMLSS amplia as ferramentas disponíveis no software R, beneficiando pesquisadores e profissionais de saúde que lidam com dados complexos. A metodologia é replicável em diferentes contextos, oferecendo instrumentos úteis para instituições de saúde, centros de pesquisa e gestores públicos no planejamento de políticas de prevenção, tratamento e acompanhamento de doenças crônicas. O trabalho impacta diretamente a comunidade científica e o setor de saúde, com potencial de beneficiar pacientes, familiares e profissionais envolvidos no cuidado, e está alinhado aos Objetivos de Desenvolvimento Sustentável da ONU como, Saúde e Bem-Estar (ODS 3), Educação de Qualidade (ODS 4) e Trabalho Decente e Crescimento Econômico (ODS 8). Assim, os impactos identificados são tanto concretos, pela aplicação prática em dados reais, quanto potenciais, pela possibilidade de extrapolação para diferentes cenários clínicos e epidemiológicos, fortalecendo avanços metodológicos e estratégicos na área da saúde.

IMPACT INDICATORS

This dissertation presents a flexible statistical model for survival analysis in patients with pancreatic cancer, one of the leading causes of morbidity and mortality worldwide. The study contributes to public health by enabling more accurate survival estimates, supporting clinical decision-making, identifying relevant prognostic factors, and promoting the improvement of patients' quality of life. From a technological perspective, the implementation of the log-normal model with a cure fraction within the GAMLSS framework expands the tools available in the R software, benefiting researchers and healthcare professionals who work with complex data. The methodology is replicable in different contexts, providing valuable instruments for healthcare institutions, research centers, and public managers in planning policies for the prevention, treatment, and monitoring of chronic diseases. The work directly impacts the scientific community and the healthcare sector, with the potential to benefit patients, their families, and professionals involved in care, and is aligned with the United Nations Sustainable Development Goals, including Good Health and Well-Being (SDG 3), Quality Education (SDG 4), and Decent Work and Economic Growth (SDG 8). Thus, the identified impacts are both concrete, through practical application to real patient data, and potential, due to the possibility of extrapolation to different clinical and epidemiological scenarios, strengthening methodological and strategic advances in the healthcare field.

LISTA DE FIGURAS

Figura 3.1 – Tipos de censura: à direita, à esquerda e intervalar.	18
Figura 3.2 – Exemplificação de mecanismos de censura à direita, em que ● falha e ○ censura.	19
Figura 3.3 – Exemplos da função $S(t)$ pelo método Kaplan-Meier.	25
Figura 3.4 – Representação de como um resíduo quantil normalizado r é obtido para uma distribuição contínua.	39
Figura 3.5 – Representação de como um resíduo quantil randomizado normalizado r é obtido para uma distribuição discreta.	40
Figura 3.6 – Falhas no ajuste adequado do parâmetro de locação do modelo, indicadas pelo <i>worm plot</i>	41
Figura 3.7 – Falhas no ajuste adequado do parâmetro de escala do modelo, indicadas pelo <i>worm plot</i>	41
Figura 3.8 – Falhas no ajuste adequado do parâmetro de assimetria do modelo, indicadas pelo <i>worm plot</i>	42
Figura 3.9 – Falhas no ajuste adequado do parâmetro de curtose do modelo, indicadas pelo <i>worm plot</i>	42

LISTA DE QUADROS

Quadro 3.1 – Tabela de contingência gerada no tempo t_a	21
Quadro 3.2 – Distribuições contínuas implementadas no pacote <code>gamlss.dist</code> , com funções de ligação padrão. (R_T = domínio de T)	32
Quadro 3.3 – Distribuições discretas implementadas no <code>gamlss.dist</code> , com funções de ligação padrão. (R_T = domínio de T , <code>inf</code> = inflado, <code>adj</code> = ajustado, <code>inv</code> = inverso)	33
Quadro 3.4 – Distribuições mistas implementadas no pacote <code>gamlss.dist</code> , com funções de ligação padrão (R_T = domínio de T).	34

SUMÁRIO

	PRIMEIRA PARTE - UM PANORAMA GERAL	12
1	INTRODUÇÃO	13
2	OBJETIVOS	16
2.1	Objetivo geral	16
2.2	Objetivos específicos	16
3	REVISÃO DE LITERATURA	17
3.1	Análise de sobrevivência	17
3.1.1	Conceitos iniciais	17
3.1.2	Kaplan-Meier	20
3.1.3	Teste Log-Rank	21
3.1.4	Modelos paramétricos clássicos	23
3.1.5	Modelos com fração de cura	24
3.2	GAMLSS	27
3.2.1	Generalização dos modelos LM, GLM e GAM	27
3.2.2	Definição matemática	30
3.2.3	Distribuições	31
3.2.4	Estimação	34
3.2.5	GAMLSS em análise de sobrevivência	35
3.2.6	Processo de seleção do modelo	36
3.2.7	Diagnóstico do modelo	38
	SEGUNDA PARTE - ARTIGO	43
	ARTIGO - Fatores prognósticos no câncer de pâncreas: uma abordagem a partir de modelos de regressão com fração de cura	43
	TERCEIRA PARTE - CONSIDERAÇÕES FINAIS	64
4	CONSIDERAÇÕES FINAIS	65
	REFERÊNCIAS	66
	APÊNDICES	69

PRIMEIRA PARTE - UM PANORAMA GERAL

1 INTRODUÇÃO

O câncer é uma doença que afeta mundialmente milhares de pessoas, representando uma das principais causas de morbidade e mortalidade, e impactando tanto a saúde pública quanto a qualidade de vida das pessoas diagnosticadas. Sendo um importante problema de saúde pública em países desenvolvidos e em desenvolvimento, o câncer corresponde à primeira ou à segunda causa de morte prematura, antes dos 70 anos (INCA, 2023).

Essa doença pode ser caracterizada pela rápida criação de células, as quais crescem além de seus limites habituais e podem invadir partes adjacentes do corpo e se espalhar para outros órgãos (Saúde, 2024). Além disso, existem mais de 100 tipos de câncer, e cada um deles pode se manifestar de maneira distinta, com características próprias de crescimento, disseminação e resposta ao tratamento.

Os sintomas do câncer são frequentemente inespecíficos e, na maioria das vezes, a doença só é detectada quando já está em uma fase avançada, o que compromete as opções de tratamento e reduz as chances de cura, como é o caso do câncer de pâncreas. Apesar da diversidade de suas causas, o desenvolvimento da doença pode ser influenciado por fatores como estilo de vida, envelhecimento da população, obesidade, infecções e, frequentemente, predisposição hereditária (INCA, 2023).

No Brasil, o câncer de pâncreas, embora não esteja entre os tipos mais frequentes, apresenta uma incidência crescente nos últimos anos e é responsável por cerca de 5% das mortes por câncer (INCA, 2023). Essa neoplasia está entre as mais desafiadoras, caracterizada por prognósticos geralmente desfavoráveis devido à detecção tardia e à rápida progressão da doença. Sua ocorrência está associada a fatores como estilo de vida, envelhecimento da população, obesidade, consumo de tabaco e, frequentemente, predisposição hereditária (INCA, 2023). Apesar do caráter altamente letal, avanços nos cuidados paliativos têm proporcionado a alguns pacientes aumento significativo na sobrevida, além de melhorias na qualidade de vida, mesmo em estágios avançados da doença.

Em estatística, várias áreas são aplicadas ao estudo do câncer, especialmente no contexto de dados que envolvem censura e fração de cura. Uma das mais relevantes é a análise de sobrevivência, que apesar de possuir aplicações em diversos campos, como economia, engenharia, sociologia e demografia, foi nas ciências biomédicas que seu desenvolvimento encontrou maior

relevância, oferecendo métodos específicos para estudar o tempo até a ocorrência de um evento de interesse, como o óbito ou a progressão da doença (Rocha; Papoila, 2009).

A análise de sobrevivência envolve um conjunto de técnicas que lidam com dados em que o evento de interesse pode não ocorrer durante o período de observação para todos os indivíduos, gerando dados censurados (Rocha; Papoila, 2009). O principal interesse é analisar o tempo do início do experimento até a ocorrência de um evento, assim, o objeto de sobrevivência é composto pelo tempo e por uma variável indicadora de censura. A censura ocorre quando, para alguns indivíduos no estudo, não é possível observar o acontecimento de interesse, como o falecimento ou a recaída da doença, dentro do tempo de acompanhamento.

Além disso, em algumas situações, é possível que uma fração dos indivíduos no estudo nunca experimente o evento de interesse, independentemente do tempo de seguimento (Colosimo; Giolo, 2024). Esse fenômeno é modelado através de modelos de fração de cura, que permitem estimar a proporção de indivíduos que, possuem uma cura ou um risco de evento nulo ao longo do tempo.

Sendo assim, uma das principais vantagens da análise de sobrevivência é que ela permite a inclusão de dados com essas características, permitindo uma análise estatística mais precisa. Para construir modelos que representem esses dados são, usualmente, utilizadas distribuições específicas, como a exponencial, Weibull, log-normal e log-logística, denominadas modelos probabilísticos ou paramétricos, que são capazes de descrever de maneira adequada diversas variáveis (Colosimo; Giolo, 2024).

Além das distribuições clássicas utilizadas na análise de sobrevivência, existem extensões desses modelos que permitem uma modelagem ainda mais precisa e adaptada às características específicas dos dados (Colosimo; Giolo, 2024). Essas extensões são fundamentais quando os dados apresentam complexidades que não podem ser totalmente capturadas pelas distribuições tradicionais.

Uma dessas extensões é a utilização dos modelos aditivos generalizados para locação, escala e forma - GAMLSS (Rigby; Stasinopoulos, 2005), atualmente também denominados como modelos de regressão distribucional (Heller; Robledo; Marschner, 2022). Os GAMLSS são modelos flexíveis que generalizam os modelos estatísticos usualmente empregados na literatura, como os modelos lineares generalizados (GLM), propostos por Nelder & Wedderburn (1972) e os modelos aditivos generalizados (GAM), apresentados por Hastie & Tibshirani (1990). Neles, a distribuição da variável dependente não se limita à família exponencial e to-

dos os parâmetros (não apenas a média) são modelados em termos de efeitos fixos e aleatórios (Stasinopoulos; Rigby, 2007). A utilização dos GAMLSS em análise de sobrevivência tem sido destacada em trabalhos como o de Castro, Cancho & Rodrigues (2010), Cordeiro, Ortega & Ramires (2015), Ramires *et al.* (2018), Ramires *et al.* (2019) e Nakamura *et al.* (2022a).

Diante do contexto apresentado, entender como o câncer se desenvolve, especialmente o câncer de pâncreas, exige o uso de métodos estatísticos capazes de lidar com dados incompletos, quando o desfecho ainda não ocorreu para todos os pacientes, e com a possibilidade de que alguns deles estejam efetivamente curados. A análise de sobrevivência, aliada aos GAMLSS, representa uma abordagem interessante para investigar de forma mais detalhada a evolução da doença e os fatores que influenciam o tempo de vida dos pacientes. Esses métodos ajudam a melhorar os estudos na área da saúde e podem apoiar ações mais eficazes no cuidado dos pacientes.

2 OBJETIVOS

Neste capítulo são apresentados o objetivo geral e os objetivos específicos que orientaram o desenvolvimento deste trabalho.

2.1 Objetivo geral

Esta dissertação tem como objetivo geral a proposição do modelo log-normal com fração de cura no contexto dos GAMLSS.

2.2 Objetivos específicos

- a) formalizar matematicamente o modelo log-normal na estrutura dos GAMLSS;
- b) implementar computacionalmente o modelo no pacote `gamlss` disponível no software R;
- c) disponibilizar os códigos necessários para sua aplicação;
- d) demonstrar a relevância do modelo por meio de sua aplicação em dados sobre câncer de pâncreas.

3 REVISÃO DE LITERATURA

Neste capítulo são discutidos os fundamentos da análise de sobrevivência, incluindo métodos clássicos e modelos com fração de cura, e a estrutura dos GAMLSS, com ênfase em sua formulação, estimação e aplicação em dados de sobrevivência.

3.1 Análise de sobrevivência

A análise de sobrevivência é uma área da estatística dedicada ao estudo do tempo até a ocorrência de um evento específico, frequentemente chamado de tempo de falha (Colosimo; Giolo, 2024). Essa metodologia é utilizada em estudos que envolvem o tempo até um desfecho particular, como a longevidade de pacientes ou a durabilidade de produtos, permitindo compreender e prever esses padrões temporais.

A caracterização de dados de sobrevivência requer o entendimento de conceitos importantes, como os tempos de sobrevida, os diferentes tipos de censura e fração de cura (Colosimo; Giolo, 2024). Esses conceitos, serão explorados a seguir.

3.1.1 Conceitos iniciais

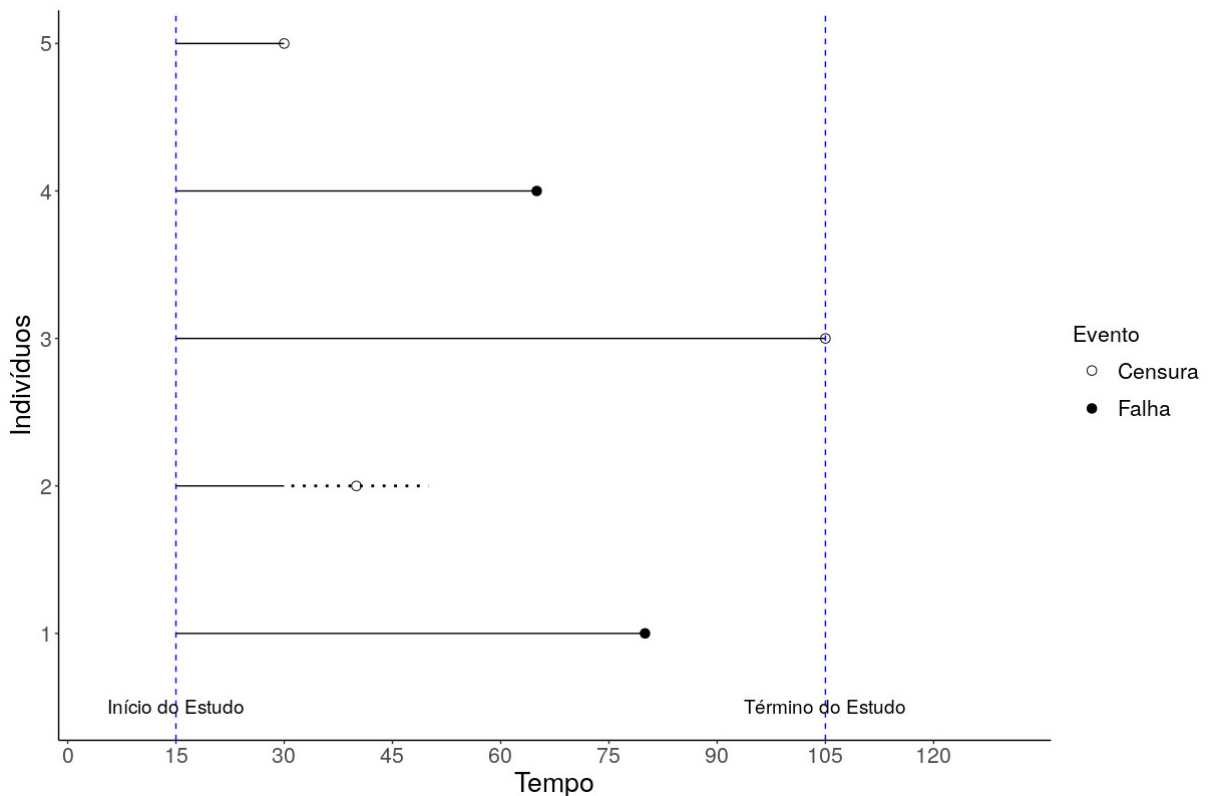
Os dados utilizados na análise de sobrevivência apresentam como característica o fato de serem longitudinais, ou seja, são informações coletadas ao longo do tempo, acompanhando indivíduos ou unidades e registrando o período até a ocorrência de um evento específico (Colosimo; Giolo, 2024).

A censura ocorre quando, por algum motivo, o evento de interesse não é registrado dentro do período estabelecido no estudo. Essa situação pode surgir, por exemplo, quando o acompanhamento é encerrado antes que todos os eventos tenham sido observados ou quando um indivíduo deixa o estudo sem que o evento tenha ocorrido. Apesar de incompletas, as observações censuradas fornecem informações importantes sobre o tempo de sobrevida dos pacientes, tornando importante sua inclusão na modelagem estatística (Colosimo; Giolo, 2024).

Os dados de sobrevivência podem apresentar diferentes tipos de censura. A censura à direita é a mais frequente e ocorre quando o paciente entra no estudo e é acompanhado por um período, mas o evento ainda não aconteceu até o final do acompanhamento ou até sua saída do

estudo. Na censura à esquerda, o paciente já apresenta o evento de interesse antes do início do acompanhamento, e sabe-se apenas que ele ocorreu em algum momento anterior ao tempo de entrada no estudo. Já na censura intervalar, o tempo exato do evento é desconhecido, mas está delimitado dentro de um intervalo de tempo (Moore, 2016). A Figura 3.1 apresenta os tipos de censura

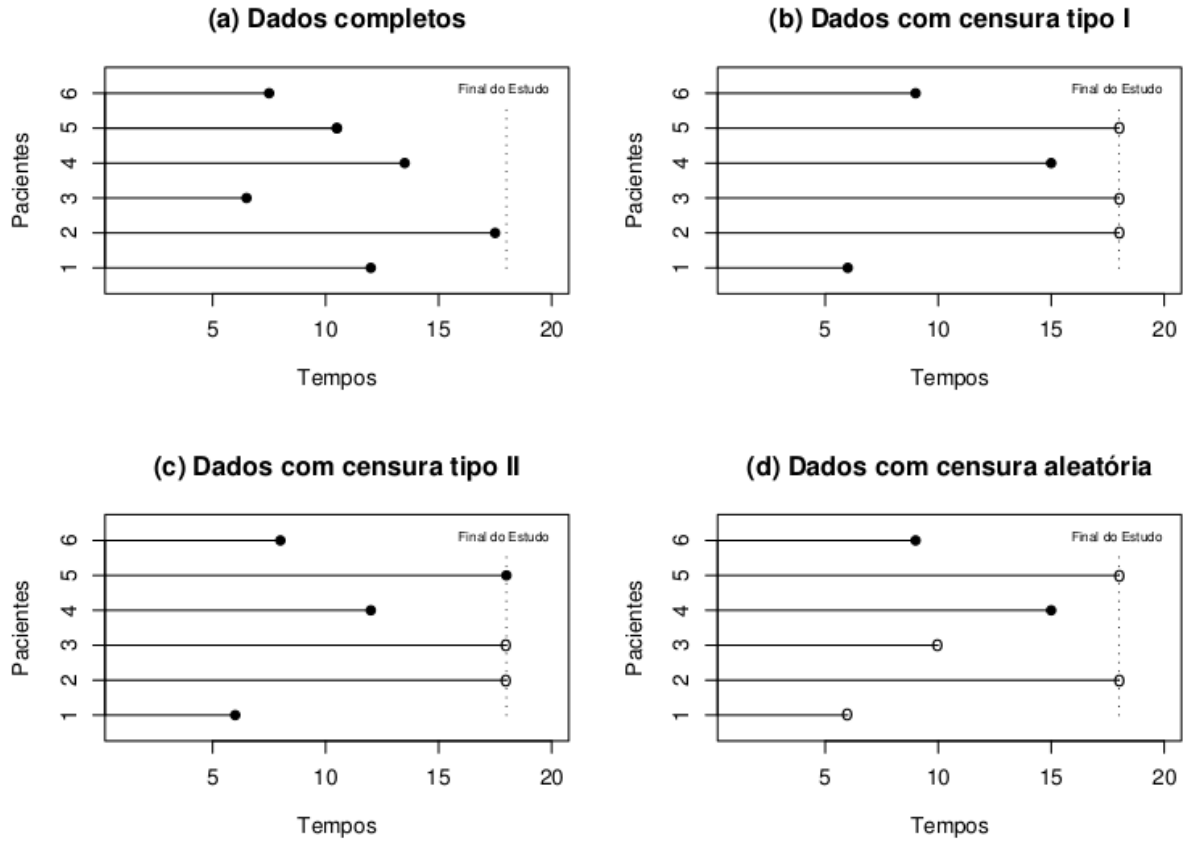
Figura 3.1 – Tipos de censura: à direita, à esquerda e intervalar.



Fonte: Do autor

Sendo assim, a censura é um elemento central na análise de sobrevivência. A censura à direita pode ser classificada em três tipos: censura do tipo I, que ocorre quando o estudo será terminado após um período pré-determinado; censura do tipo II, ocorre quando o estudo é encerrado assim que o evento de interesse acontece em um número previamente definido de indivíduos; e censura aleatória, na qual os tempos de censura variam de forma imprevisível entre as observações (Moore, 2016). A Figura 3.2, mostra os tipos de censura à direita.

Figura 3.2 – Exemplificação de mecanismos de censura à direita, em que ● falha e ○ censura.



Fonte: Colosimo & Giolo (2024)

De acordo com Colosimo & Giolo (2024), podemos representar o mecanismo da censura aleatória da seguinte maneira

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

em que, t é o valor observado; T é uma variável aleatória que representa o tempo de falha; C é uma variável aleatória independente de T que representa o tempo de censura associado ao indivíduo de estudo.

A função de sobrevivência é utilizada para modelar o tempo de falha, tempo de sobrevivência ou tempo de evento (Harrell, 2015). Denota-se a variável resposta como T , representando o tempo até o evento. Em vez de definir o modelo estatístico com base no tempo esperado de falha, é mais vantajoso fazê-lo em termos da função de sobrevivência, $S(t)$. Esta função é definida como a probabilidade de uma observação não falhar até o tempo t , ou seja, a probabi-

lidade de uma observação sobreviver até o tempo t . Em termos probabilísticos, isso é expresso como

$$S(t) = P(T > t), \quad t \geq 0. \quad (3.1)$$

Consequentemente, a função de distribuição acumulada, representada por $F(t) = 1 - S(t)$, é definida como a probabilidade de uma observação falhar até o tempo t , ou seja, a probabilidade de que o evento de interesse ocorra antes ou exatamente no tempo t .

De acordo com Colosimo & Giolo (2024), a função taxa de falha é bastante útil para descrever a distribuição do tempo de sobrevida de pacientes, já que ela descreve a forma em que a taxa instantânea de falha se altera com o tempo, e pode ser definida por

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

Outra função importante é a taxa de falha acumulada, que não tem uma interpretação direta, mas pode ser útil na avaliação da função de maior interesse que é a de taxa de falha, $h(t)$, (Colosimo; Giolo, 2024). Ela pode ser definida como

$$H(t) = \int_0^t h(u) du.$$

Com isso, surge a necessidade de estimar a função de sobrevivência, e o estimador de Kaplan-Meier é um dos mais utilizados para essa finalidade.

3.1.2 Kaplan-Meier

O estimador não-paramétrico de Kaplan-Meier, proposto por Kaplan & Meier (1958), permite calcular a função de sobrevivência de maneira não paramétrica, ajustando-se adequadamente aos dados censurados e proporcionando uma estimativa precisa do tempo de sobrevivência dos indivíduos.

Ele é definido como o produto, ao longo dos tempos de falha, das probabilidades condicionais de sobrevivência até o próximo tempo de falha (Moore, 2016). Conhecido como estimador de produto-limite, ele representa o caso limite das estimativas de sobrevivência atuariais, à medida que os intervalos de tempo se tornam menores, permitindo a inclusão de cada tempo de falha individualmente (Harrell, 2015). Formalmente, é definido como

$$\hat{S}(t) = \prod_{t_n \leq t} \left(1 - \frac{d_a}{n_a}\right),$$

em que $t_1 < t_2 < \dots < t_b$ os b tempos de falhas distintos e ordenados de falha, d_b o total de falhas em $(t_{a-1}, t_a]$ e n_a o total de indivíduos sob risco em t_{a-1} , $a = 1, \dots, b$.

Esse estimador possui a característica de ser não viciado para amostras grandes, ser fracamente consistente, convergir assintoticamente para um processo gaussiano e além disso, ser o estimador de máxima verossimilhança de $S(t)$ (Colosimo; Giolo, 2024). Assim, com o estimador de Kaplan-Meier fornecendo uma estimativa da função de sobrevivência, é possível comparar as curvas de sobrevivência entre diferentes grupos.

3.1.3 Teste Log-Rank

O teste Log-Rank é utilizado para comparar duas funções de sobrevivência $S_1(t)$ e $S_2(t)$ ao longo do tempo, avaliando se há diferenças significativas entre elas. Para isso, consideram-se os tempos de falha distintos $t_1 < t_2 < \dots < t_b$, obtidos pela combinação das duas amostras. Em cada tempo t_a , os dados podem ser organizados em uma tabela de contingência 2×2 , conforme apresentado na Quadro 3.1 (Colosimo; Giolo, 2024).

Quadro 3.1 – Tabela de contingência gerada no tempo t_a

	Grupos		Totais
	1	2	
Falha	d_{1a}	d_{2a}	d_a
Não Falha	$n_{1a} - d_{1a}$	$n_{2a} - d_{2a}$	$n_a - d_a$
Totais	n_{1a}	n_{2a}	n_a

Fonte: Colosimo & Giolo (2024)

Suponha que no tempo t_a , ocorrem d_a falhas e n_a indivíduos estão sob risco no instante imediatamente anterior a t_a na amostra combinada. Para cada grupo $o = 1, 2$ e $a = 1, 2, \dots, b$, temos d_{ia} falhas e $n_{ia} - d_{ia}$ sobreviventes. A tabela é estruturada de forma que as marginais representam as somas por linhas e colunas, como descrito por Moore (2016). Por exemplo, o total de falhas em t_a é dado por $d_a = d_{1a} + d_{2a}$, e o número total de indivíduos em risco é $n_a = n_{1a} + n_{2a}$.

Sob a hipótese nula de igualdade das funções de sobrevivência, d_{ia} segue uma distribuição hipergeométrica, como destacado por Colosimo & Giolo (2024) e Moore (2016), com função de probabilidade dada por

$$P(d_{1a}|n_{1a}, n_{2a}, d_a) = \frac{\binom{n_{1a}}{d_{1a}} \binom{n_{2a}}{d_{2a}}}{\binom{n_a}{d_a}}.$$

Esse modelo reflete a alocação das falhas d_a entre os grupos proporcionalmente ao número de indivíduos sob risco em cada grupo, na ausência de diferenças entre as populações.

A média esperada de d_{2a} (ou $E(d_{2a})$) é dada por

$$w_{2a} = \frac{n_{2a}d_a}{n_a}, \quad (3.2)$$

e sua variância por

$$(V_a)_2 = \frac{n_{2a}(n_a - n_{2a})d_j(n_a - d_a)}{n_a^2(n_a - 1)}. \quad (3.3)$$

As expressões 3.2 e 3.3, respectivamente, quantificam a expectativa e a variabilidade do número de falhas no grupo 2, considerando a estrutura condicional dos dados. A estatística de teste log-rank, que sintetiza a discrepância entre as falhas observadas e esperadas ao longo de b tempos de falha, é definida como

$$T = \frac{[\sum_{a=1}^b (d_{2a} - w_{2a})]^2}{\sum_{a=1}^b (V_a)_2} \sim \chi_{(1)}^2.$$

Sob a hipótese nula de igualdade entre as funções de sobrevivência, T segue uma distribuição qui-quadrado com 1 grau de liberdade para amostras grandes. Assim, o teste avalia se as diferenças observadas são compatíveis com a variabilidade esperada pela distribuição hipergeométrica, fornecendo uma base robusta para inferências sobre a igualdade das funções de sobrevivência entre dois grupos.

3.1.4 Modelos paramétricos clássicos

Conforme Colosimo & Giolo (2024), diversos modelos probabilísticos ou paramétricos foram propostos na literatura para descrever o tempo até a falha. Dentre eles, os modelos Exponencial, Weibull e log-normal se destacam por sua aplicabilidade e adequação a uma variedade de problemas práticos.

O modelo exponencial é um dos modelos paramétricos mais simples empregados para descrever o tempo até a ocorrência de um evento. De acordo com Colosimo & Giolo (2024), esse modelo apresenta apenas um parâmetro e assume que a taxa de risco é constante ao longo do tempo, o que significa que a probabilidade de ocorrência do evento de interesse não varia com o tempo. Se o tempo de falha, representado pela variável aleatória T , segue uma distribuição exponencial, então a função densidade de probabilidade é dada pela seguinte expressão

$$f(t) = \frac{1}{\alpha} \exp \left[- \left(\frac{t}{\alpha} \right) \right], \quad t \geq 0,$$

em que o parâmetro $\alpha > 0$ representa o tempo médio de vida. As funções de sobrevivência e de taxa de falha são dadas, respectivamente, por

$$\begin{aligned} S(t) &= \exp \left[- \left(\frac{t}{\alpha} \right) \right], \quad \text{para } t \geq 0, \\ h(t) &= \frac{1}{\alpha}, \quad \text{para } t \geq 0. \end{aligned} \tag{3.4}$$

A distribuição de Weibull (Weibull, 1939), é uma das distribuições mais utilizadas. Segundo Colosimo & Giolo (2024), sua popularidade se explica pela capacidade de apresentar uma ampla variedade de formas e pela propriedade de sua função de taxa de falha ser monótona, podendo ser crescente, decrescente ou constante. Nesse contexto, a função densidade de probabilidade é descrita como

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp \left[- \left(\frac{t}{\alpha} \right)^\gamma \right], \quad \text{para } t \geq 0,$$

com os parâmetros de escala ($\alpha > 0$) e forma ($\gamma > 0$). As funções de sobrevivência e de taxa de falha são dadas, respectivamente, por

$$\begin{aligned} S(t) &= \exp \left[- \left(\frac{t}{\alpha} \right)^\gamma \right], \quad t \geq 0, \\ h(t) &= \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}, \quad t \geq 0. \end{aligned} \quad (3.5)$$

Em termos matemáticos, no modelo log-normal, se o logaritmo do tempo de sobrevivência segue uma distribuição normal com média $\mu \in \mathbb{R}$ e desvio-padrão $\sigma > 0$, então o tempo de sobrevivência T segue uma distribuição log-normal (Colosimo; Giolo, 2024).

Assim, para uma variável aleatória T que segue uma distribuição log-normal, a função densidade de probabilidade é dada por

$$f(t) = \frac{1}{\sqrt{2\pi t} \sigma} \exp \left[-\frac{1}{2} \left(\frac{\log(t) - \mu}{\sigma} \right)^2 \right], \quad \text{para } t \geq 0.$$

As funções de sobrevivência e taxa de falha não apresentam uma forma analítica fechada (Colosimo; Giolo, 2024) e podem ser determinadas, respectivamente, como

$$\begin{aligned} S(t) &= \Phi \left(\frac{-\log(t) + \mu}{\sigma} \right) \\ h(t) &= \frac{f(t)}{S(t)}, \end{aligned} \quad (3.6)$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão.

3.1.5 Modelos com fração de cura

Na análise de sobrevivência, usualmente, assume-se que todos os indivíduos sob estudo estão sujeitos à ocorrência do evento de interesse e que, se acompanhados por tempo suficiente, o evento será observado. No entanto, esse pressuposto pode não ser razoável em determinadas situações, pois uma fração da população pode não estar suscetível a experimentar o evento, mesmo em períodos de acompanhamento muito longos (Colosimo; Giolo, 2024).

Para lidar com esse tipo de cenário, onde uma parte dos indivíduos não experimentam o evento, surge o modelo de fração de cura. Esse modelo considera os indivíduos que, devido a

alterações em certos fatores ou à inexistência de fatores de risco, não experimentarão o evento de interesse ao longo do período de acompanhamento. Assim, o modelo de fração de cura, busca estimar a fração de indivíduos curados, oferecendo uma abordagem mais precisa em estudos onde a ocorrência do evento é rara ou de baixa probabilidade (Pal *et al.*, 2023).

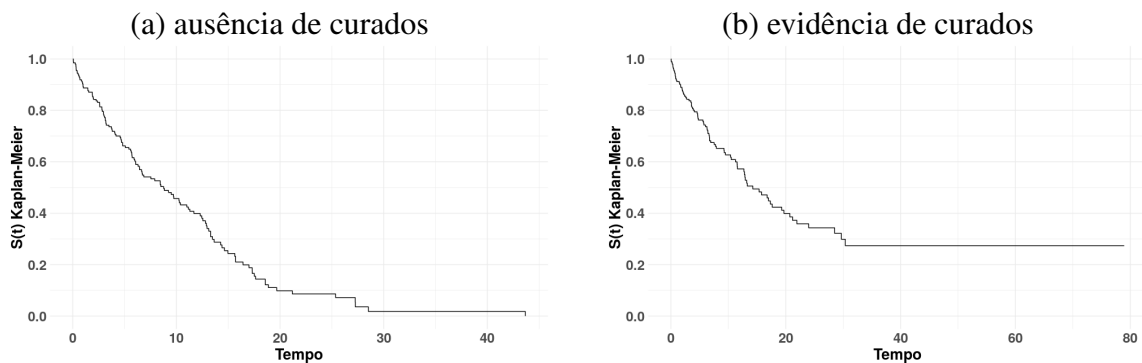
Quando há indivíduos imunes ou curados, a função de sobrevivência $S(t)$ caracteriza-se por uma diminuição ao longo do tempo, sem atingir o valor zero, mesmo que o tempo de seguimento t tenda ao infinito, ou, equivalentemente, a função de distribuição $F(t) = 1 - S(t)$ cresce, mas não atinge o valor 1, quando t tende ao infinito (Colosimo; Giolo, 2024), ou seja

$$\lim_{t \rightarrow \infty} S(t) = (1 - \pi) > 0 \quad \text{ou} \quad \lim_{t \rightarrow \infty} F(t) = \pi < 1,$$

em que $(1 - \pi)$ é a probabilidade de ser suscetível ao evento, π a probabilidade de não ser suscetível ao evento. Como $S(t)$ não atinge o valor zero e $F(t)$ não chega a 1, a função $S(t)$ é chamada de função de sobrevivência imprópria, enquanto $F(t)$ é considerada uma função de distribuição imprópria (Colosimo; Giolo, 2024).

Uma forma de verificar a presença de indivíduos curados na população foi proposta por Maller & Zhou (1996), que sugere estimar a função $S(t)$ utilizando o estimador não paramétrico de Kaplan-Meier e, com base na curva estimada $\hat{S}(t)$, avalia-se se ela é imprópria, ou seja, se $\lim_{t \rightarrow \infty} \hat{S}(t) > 0$, também avalia-se se o período de tempo é razoável. A Figura 3.3 ilustra o comportamento da função de sobrevivência $S(t)$ estimada.

Figura 3.3 – Exemplos da função $S(t)$ pelo método Kaplan-Meier.



Fonte: Do autor

Nesse contexto, os modelos de mistura padrão são modelos com fração de cura muito populares para realizar uma análise do tempo de sobrevida. Esse modelo foi inicialmente proposto por (Boag, 1949), e, posteriormente, retomado por (Berkson; Gage, 1952). Ele assume a

existência de uma fração $(1 - \pi)$ de indivíduos não curados e, uma proporção π de indivíduos curados. Assim, considerando U uma variável aleatória Bernoulli, em que $U = 1$ o indivíduo é suscetível ao evento, e $U = 0$, caso contrário, a função de sobrevivência imprópria associada à população de interesse, com $t > 0$ é dada por

$$S_{pop}(t) = \pi + (1 - \pi)S_U(t) \quad (3.7)$$

em que $(1 - \pi)$ é a probabilidade de ser suscetível ao evento, π a probabilidade de não ser suscetível ao evento e $S_U(t)$ é a função de sobrevivência própria associada aos indivíduos não imunes. Quando não houver curados ($\pi = 1$), o modelo de mistura se reduz ao modelo de sobrevivência padrão, apresentado na equação 3.1.

Na presença de covariáveis, o modelo 3.7 fica descrito como

$$S_{pop}(t|\mathbf{x}, \mathbf{z}) = [1 - \pi(\mathbf{z})] + \pi(\mathbf{z})S_U(t|\mathbf{x}), \quad t \geq 0,$$

em que $\mathbf{z}$ é o vetor de covariáveis de dimensão q que exerce efeito sobre $\pi(\mathbf{z}) = P(U = 1|\mathbf{z})$ e $\mathbf{x}$ é o vetor de covariáveis de dimensão p que exerce efeito sobre $S_U(t|\mathbf{x})$ que é a função de sobrevivência própria associada aos indivíduos não imunes (Colosimo; Giolo, 2024). Com isso, a função densidade imprópria é dada por

$$f_{pop}(t | \mathbf{x}, \mathbf{z}) = \pi(\mathbf{z})f_U(t | \mathbf{x}),$$

em que $f_U(t|\mathbf{x})$ corresponde a função densidade própria associada ao grupo dos indivíduos que estão sob risco. E a função taxa de falha populacional pode ser expressa por

$$h_{pop}(t | \mathbf{x}, \mathbf{z}) = \frac{f_{pop}(t | \mathbf{x}, \mathbf{z})}{S_{pop}(t | \mathbf{x}, \mathbf{z})}.$$

Para diferentes distribuições paramétricas, a combinação da função de sobrevivência da população com fração de cura (Equação 3.7) e a função de sobrevivência específica da distribuição em questão resulta em modelos de sobrevivência que incorporam a presença de indivíduos não suscetíveis ao evento.

No caso do modelo exponencial, ao combinar as funções de sobrevivência 3.4 e 3.7, obtém-se o modelo de sobrevivência exponencial com fração de cura

$$S_{pop}(t) = \pi + (1 - \pi) \left[-\left(\frac{t}{\alpha} \right) \right]. \quad (3.8)$$

A função densidade de probabilidade associada a 3.8 é dada por

$$f_{pop}(t) = (1 - \pi) \frac{1}{\alpha} \exp \left[- \left(\frac{t}{\alpha} \right) \right].$$

Para o modelo Weibull, a combinação das Equações 3.5 e 3.7 conduz ao modelo de sobrevivência Weibull com fração de cura

$$S_{pop}(t) = \pi + (1 - \pi) \exp \left[- \left(\frac{t}{\alpha} \right)^\gamma \right]. \quad (3.9)$$

Do mesmo modo, a função densidade de probabilidade associada a 3.9 é definida como

$$f_{pop}(t) = (1 - \pi) \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp \left[- \left(\frac{t}{\alpha} \right)^\gamma \right].$$

Por fim, para o modelo log-normal, combinando as funções de sobrevivência 3.6 e 3.7, o modelo de sobrevivência log-normal com fração de cura é dado por

$$S_{pop}(t) = \pi + (1 - \pi) \Phi \left(\frac{-\log(t) + \mu}{\sigma} \right). \quad (3.10)$$

E a função densidade de probabilidade associada a 3.10 é expressa por

$$f_{pop}(t) = (1 - \pi) \frac{1}{\sqrt{2\pi}t\sigma} \exp \left[-\frac{1}{2} \left(\frac{\log(t) - \mu}{\sigma} \right)^2 \right].$$

3.2 GAMLSS

Nesta seção, será abordada a trajetória da modelagem estatística, partindo dos modelos lineares (LM), passando pelos modelos lineares generalizados (GLM) e aditivos generalizados (GAM), até chegar ao modelo aditivo generalizado para locação, escala e forma (GAMLSS). Serão exploradas as distribuições disponíveis no GAMLSS, os métodos para estimar seus parâmetros, as estratégias de diagnóstico do modelo e suas aplicações em análise de sobrevivência.

3.2.1 Generalização dos modelos LM, GLM e GAM

A regressão linear é uma técnica utilizada para estimar o valor de uma variável de interesse (variável resposta) com base em uma ou mais variáveis explicativas. Nesse modelo,

assume-se que a variável resposta segue uma distribuição normal e que sua média está diretamente relacionada a uma combinação linear dos preditores, utilizando uma função identidade para estabelecer essa conexão. De acordo com Stasinopoulos *et al.* (2017), considerando T a variável resposta, com r covariáveis $x_1, \dots, x_r$ e tamanho amostral n , o modelo pode ser expresso como

$$T_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_r x_{ir} + \varepsilon_i,$$

sendo $\varepsilon_i \stackrel{ind}{\sim} N(0, \sigma^2)$ para $i = 1, 2, \dots, n$. Este modelo pressupõe que os termos de erro ε_i são independentes e seguem uma distribuição normal com média zero e variância constante σ^2 . Pode-se dizer então que

$$\begin{aligned} T_i &\stackrel{ind}{\sim} N(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_{i1} + \dots + \beta_r x_{ir} \text{ para } i = 1, 2, \dots, n. \end{aligned}$$

Além disso, o modelo também pode ser reescrito na forma matricial

$$\begin{aligned} \mathbf{T} &\stackrel{ind}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \\ \boldsymbol{\mu} &= \mathbf{X}\boldsymbol{\beta}, \end{aligned}$$

em que o vetor de respostas é representado por $\mathbf{T} = (T_1, \dots, T_n)^\top$, enquanto $\mathbf{X}$ denota a matriz de delineamento de dimensão $n \times p$, contendo r colunas correspondentes às covariáveis e uma coluna adicional composta exclusivamente por uns, totalizando ($p = r + 1$). O vetor de coeficientes é dado por $\boldsymbol{\beta} = (\beta_0, \dots, \beta_r)^\top$, o vetor de médias é expresso como $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^\top$, e as variâncias, assumidas constantes, são descritas por $\boldsymbol{\sigma}^2 = (\sigma^2, \dots, \sigma^2)$. A estimativa de $\boldsymbol{\beta}$ pode ser obtida aplicando-se o método dos mínimos quadrados, que minimiza a soma dos quadrados das diferenças entre as observações T_i e as respectivas médias, calculadas como $\mu_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_r x_{ir}$ (Stasinopoulos *et al.*, 2017).

A regressão linear assume que os resíduos seguem uma distribuição normal, caso essa suposição não seja atendida, os testes de hipóteses e os intervalos de confiança podem se tornar imprecisos, o que pode comprometer a validade dos resultados e exigir o uso de outros modelos.

Nesse sentido, os Modelos Lineares Generalizados (GLM), introduzidos por Nelder & Wedderburn (1972), foram desenvolvidos para ampliar a aplicação dos modelos lineares clássicos ao lidar com variáveis dependentes que não seguem a distribuição normal. Esses modelos

permitem que a variável resposta siga uma distribuição pertencente à família exponencial, definida como

$$f(t | \mu, \phi) = \exp \left(\frac{t\theta - d(\theta)}{\phi} + c(t, \phi) \right), \quad (3.11)$$

em que $\mathbb{E}(T) = \mu = d'(\theta)$ e $\text{Var}(Y) = \phi V(\mu)$, com $V(\mu) = d''(\theta(\mu))$ chamada de função de variância. A estrutura da equação 3.11 engloba diversas distribuições relevantes, como a binomial, Poisson ou gama, tornando-os adequados para uma ampla variedade de situações.

De acordo com Casella & Berger (2001), um GLM é composto por três elementos principais: o componente aleatório, o componente sistemático e a função de ligação. O componente aleatório descreve a distribuição da variável resposta T_i , assumindo que ela pertence à família exponencial de distribuições, com observações independentes entre si. Já o componente sistemático é uma combinação linear dos preditores x_{ij} , expressa como $\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$, e reflete a relação entre as variáveis preditoras e a média da variável resposta. Por fim, a função de ligação estabelece a conexão entre a média da variável resposta, $\mu_i = E[T_i]$, e o preditor linear η_i , sendo geralmente representada por $g(\mu_i) = \eta_i$. A inversa da função de ligação, $g^{-1}(\eta_i)$, permite retornar à escala original da média.

Formalmente, um GLM pode ser descrito como

$$\begin{aligned} T_i &\sim \varepsilon(\mu_i, \phi) \\ g(\mu_i) = \eta_i &\iff \mu_i = g^{-1}(\eta_i), \end{aligned}$$

em que T_i , $i = 1, \dots, n$, é a variável resposta que segue uma distribuição da família exponencial com média μ_i e parâmetro de dispersão ϕ .

Os GLM ampliam a flexibilidade dos modelos de regressão linear (LM) ao dispensarem o pressuposto de normalidade, mas ainda enfrentam a limitação de assumir uma relação linear entre a média da variável resposta e as variáveis explicativas. Para superar essa restrição, Hastie & Tibshirani (1990) introduziram as funções de suavização nos GLMs, dando origem a uma nova classe de modelos conhecidos como Modelos Aditivos Generalizados (GAM). Matematicamente, um GAM é expresso por

$$\begin{aligned} \mathbf{T} &\overset{ind}{\sim} \varepsilon(\boldsymbol{\mu}, \boldsymbol{\phi}) \\ \boldsymbol{\eta} = g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + s_1(\mathbf{x}_1) + \dots + s_J(\mathbf{x}_J), \end{aligned}$$

em que s_j , $j = 1, \dots, J$ é uma função de suavização não paramétrica aplicada à covariável x_j (Stasinopoulos *et al.*, 2017). Esse modelo permite que a relação entre o preditor $\eta = g(\mu)$ e

as variáveis explicativas seja determinada pelos próprios dados, sem a necessidade de impor uma forma específica, como linear ou polinomial. Ao utilizar essa abordagem as funções suavizadoras capturam os padrões de modo a se ajustar às características naturais do conjunto de dados.

Os GAM possuem algumas limitações importantes, como modelar apenas o parâmetro da média em função das covariáveis, deixando variações em outros parâmetros, como dispersão ou assimetria, sem modelagem explícita, além de que a distribuição da variável resposta pertença à família exponencial, o que reduz sua aplicabilidade em contextos mais complexos.

Diante disso, em determinados cenários, pode ser necessário adotar abordagens ainda mais adaptáveis, capazes de modelar de forma mais precisa as características específicas da distribuição da variável resposta.

3.2.2 Definição matemática

Diante do exposto, Rigby & Stasinopoulos (2005) introduziram os modelos aditivos generalizados para locação, escala e forma (GAMLSS), que se destacam por sua natureza semi-paramétrica e maior flexibilidade, representando uma generalização dos modelos apresentados anteriormente. Os GAMLSS possibilitam a modelagem não apenas da média da distribuição, mas também dos parâmetros de locação, escala e forma da variável resposta.

Nesse modelo, qualquer distribuição para a variável resposta pode ser utilizada, com todos os parâmetros da distribuição podendo ser modelados como funções das variáveis explicativas, por meio de funções lineares e/ou de suavização não-paramétricas (Stasinopoulos; Rigby; Bastiani, 2018). Assim, cada parâmetro tem sua estrutura de regressão, adaptando-se às propriedades da distribuição selecionada e às características dos dados analisados.

Um GAMLSS é definido por Rigby & Stasinopoulos (2005) como

$$T_i \sim \mathcal{D}(\boldsymbol{\theta}_k)$$

$$g_k(\boldsymbol{\theta}_k) = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} s_{jk}(\mathbf{x}_{jk}), \quad (3.12)$$

em que $T_1, T_2, \dots, T_n$ é uma amostra aleatória com $i = 1, \dots, n$, e $\mathcal{D}$ denota qualquer distribuição com vetor de parâmetros $\boldsymbol{\theta}_k$, sendo $\boldsymbol{\theta}_k = (\theta_1, \theta_2, \dots, \theta_p)^\top$. Para $k = 1, 2, \dots, p$, considera-se $g_k(\cdot)$ uma função de ligação conhecida e monótona que relaciona o parâmetro $\boldsymbol{\theta}_k$ às variáveis

explicativas e efeitos aleatórios através de um modelo aditivo, a matriz $\mathbf{X}_k$ representa a matriz de delineamento que inclui os termos lineares aditivos do modelo, enquanto o vetor $\boldsymbol{\beta}_k$ refere-se ao vetor dos parâmetros dos coeficientes lineares e s_{jk} são funções de suavização não-paramétricas para variáveis explicativas $\mathbf{x}_{jk}$.

No GAMLSS, as variáveis explicativas quantitativas, podem ser as mesmas ou diferentes das variáveis $\mathbf{x}_{kj}$ utilizadas nos suavizadores (Stasinopoulos *et al.*, 2017). Grande parte das funções suaves empregadas no modelo podem ser expressas na forma $s(\mathbf{x}) = \mathbf{Z}\boldsymbol{\gamma}$, em que $\mathbf{Z}$ é a matriz base que depende da variável explicativa $\mathbf{x}$. Aqui, $\boldsymbol{\gamma}$ é um vetor de parâmetros que precisa ser estimado, e sua estimativa é sujeita a uma penalidade quadrática da forma $\lambda \boldsymbol{\gamma}^\top \mathbf{G}\boldsymbol{\gamma}$, em que $\mathbf{G} = \mathbf{D}^\top \mathbf{D}$ é uma matriz conhecida, λ é o hiperparâmetro que controla a quantidade de suavização aplicada no ajuste e $\mathbf{D}$ é uma matriz de diferenças. Essas funções são chamadas de funções suaves penalizadas ou suavizadores penalizados.

3.2.3 Distribuições

No GAMLSS, é possível utilizar uma variedade de distribuições paramétricas para modelar a variável resposta, sem restrições quanto à sua natureza. Essas distribuições são classificadas em três grandes grupos: discretas, contínuas e mistas. Cada grupo inclui distribuições com características específicas, que permitem lidar com diferentes tipos de dados, como aqueles com assimetria acentuada e diferentes graus de curtose. Além disso, elas são adequadas para situações de heterogeneidade, em que a escala ou a forma da distribuição da variável resposta varia conforme as variáveis explicativas (Stasinopoulos; Rigby, 2007).

Nesse contexto, o pacote `gamlss` (Stasinopoulos; Rigby, 2007), disponibilizado no software R Core Team (2024), conta com mais de 100 distribuições implementadas. Além das opções já disponíveis, existe a possibilidade de criação de novas distribuições e a geração de outras por meio de transformações das existentes, oferecendo maior flexibilidade para a modelagem dos dados.

Uma revisão abrangente de todas as distribuições disponíveis nesse pacote pode ser encontrada no livro *Distributions for Modelling Location, Scale, and Shape: Using GAMLSS in R* (Rigby *et al.*, 2019). Os Quadros 3.2, 3.3 e 3.4 mostram, respectivamente, exemplos de distribuições contínuas, discretas e mistas disponíveis no pacote `gamlss.dist`.

Quadro 3.2 – Distribuições contínuas implementadas no pacote `gamlss.dist`, com funções de ligação padrão. ($R_T = \text{domínio de } T$)

Distribuição	Nome no gamlss	Função de ligação dos parâmetros			
		μ	σ	ν	τ
$R_T = (-\infty, \infty)$					
exponencial Gaussiana	exGAUS	ident	log	log	-
exponencial beta generalizada tipo 2	EGB2	ident	log	log	log
t generalizado	GT	ident	log	log	-
Gumbel	GU	ident	log	-	-
SU original de Johnson	JSUo	ident	log	ident	log
SU reparado de Johnson	JSU	ident	log	ident	log
logística	LO	ident	log	-	-
normal	NO	ident	log	-	-
exponencial potência 2	PE2	ident	log	-	-
Gumbel reverso	RG	ident	log	-	-
sinh-arcsinh	SHASH	ident	log	log	log
normal assimétrico tipo 1	SN1	ident	log	-	-
exponencial potência assimétrica tipo 1	SEP1	ident	log	log	log
t assimétrico tipo 1	ST1	ident	log	log	-
família t	TF	ident	log	-	-
$R_T = (0, \infty)$					
Box-Cox Cole Green	BCCG	ident	log	ident	-
exponencial potência Box-Cox	BCPE	ident	log	ident	log
Box-Cox t	BCT	ident	log	ident	log
exponencial	EXP	log	-	-	-
gamma	GA	log	log	-	-
beta generalizada tipo 2	GB2	log	log	log	log
gamma generalizada	GG	log	log	ident	-
inversa generalizada Gaussiana	GIG	log	log	ident	-
$R_T = (0, 1)$					
beta	BE	logit	-	-	-
beta generalizada tipo 1	GB1	logit	logit	log	log

Fonte: Stasinopoulos *et al.* (2017)

Quadro 3.3 – Distribuições discretas implementadas no `gamlss.dist`, com funções de ligação padrão. (R_T = domínio de T , inf = inflado, adj = ajustado, inv = inverso)

Distribuição	Nome no gamlss	Função de ligação dos parâmetros			
		μ	σ	ν	τ
$R_T = \{0, 1, 2, \dots\}$					
beta negativa binomial	BNB	log	log	log	-
Delaporte	DEL	log	log	logit	-
Poisson generalizada	GPO	log	-	-	-
geométrica	GEOM	log	-	-	-
binomial negativa tipo 1	NBI	log	log	-	-
Poisson	PO	log	-	-	-
Poisson-inversa Gaussiana	PIG	log	-	-	-
beta negativa binomial com ajuste de zero	ZABNB	logit	log	logit	-
logarítmica com ajuste de zero	ZALG	logit	logit	-	-
binomial negativa com ajuste de zero	ZANBI	log	log	log	-
Poisson com ajuste de zero	ZAP	logit	-	-	-
binomial negativa com inflação de zero	ZINB	log	log	logit	-
Poisson com inflação de zero	ZIP	logit	-	-	-
$R_T = \{0, 1, \dots, n\}$					
binomial	BI	logit	-	-	-
beta binomial	BB	logit	log	-	-
binomial dupla	DBI	log	log	-	-
beta binomial com ajuste de zero	ZABB	logit	logit	-	-
binomial com ajuste de zero	ZABI	logit	logit	-	-
beta binomial com inflação de zero	ZIBB	logit	log	-	-
binomial com inflação de zero	ZIBI	logit	-	-	-
$R_T = \{1, 2, \dots\}$					
logarítmica	LG	logit	-	-	-

Fonte: Stasinopoulos *et al.* (2017)

Quadro 3.4 – Distribuições mistas implementadas no pacote `gamlss.dist`, com funções de ligação padrão (R_T = domínio de T).

Distribuição	Nome no <code>gamlss</code>	R_T	Função de ligação dos parâmetros			
			μ	σ	ν	τ
beta com inflação de zero	BEINF0	$[0, 1]$	logit	logit	log	-
beta com inflação de um	BEINF1	$(0, 1]$	logit	logit	log	-
beta com inflação de zero e um	BEINF	$[0, 1]$	logit	logit	log	log
gamma com ajuste de zero	ZAGA	$[0, \infty)$	log	log	log	-
Gaussiana inversa com ajuste de zero	ZAIG	$[0, \infty)$	log	log	log	-

Fonte: Stasinopoulos *et al.* (2017)

3.2.4 Estimação

A estimação dos parâmetros no GAMLSS é realizada por meio da maximização da função de verossimilhança ou da verossimilhança penalizada, dependendo da presença de funções de suavização ou efeitos aleatórios no modelo. No caso de modelos puramente paramétricos, sem suavização ou efeitos aleatórios, apenas os parâmetros de regressão β , associados às variáveis explicativas, precisam ser estimados. Desse modo, conforme discutido por Stasinopoulos *et al.* (2017), tem-se

$$\ell = \sum_{i=1}^n \log f(t_i | \mu_i, \sigma_i, \nu_i, \tau_i). \quad (3.13)$$

Por outro lado, o modelo mais abrangente, que inclui efeitos aleatórios, é ajustado utilizando a estimativa de máxima verossimilhança penalizada, exigindo também a estimação dos parâmetros γ , que representam os coeficientes das funções de suavização, e λ , o hiperparâmetro que controla o nível de suavização. Assim, o logaritmo da função de verossimilhança penalizada é dada por

$$\ell_p = \ell - \frac{1}{2} \sum_{k=1}^4 \sum_{j=1}^{J_k} \gamma_{kj}^\top \mathbf{G}_{kj}(\lambda_{kj}) \gamma_{kj}, \quad (3.14)$$

em que $\ell = \sum_{i=1}^n \log \left\{ f(t_i | \theta^i) \right\}$ é a função de verossimilhança logaritmica, para θ^i com $i = 1, 2, \dots, n$.

Rigby & Stasinopoulos (2005) propõem os algoritmos CG e RS para ajustar um GAMLSS com hiperparâmetros fixos, com o objetivo de maximizar a função de verossimilhança penalizada. O algoritmo RS é uma generalização do método utilizado por Rigby & Stasinopoulos (1996) para ajustar modelos aditivos de média e dispersão (MADAM). Ele não faz uso das derivadas cruzadas da log-verossimilhança. Já o algoritmo CG, que é uma generalização do método de Cole & Green (1992), exige informações sobre as primeiras e segundas derivadas cruzadas da log-verossimilhança em relação aos parâmetros de distribuição $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$.

3.2.5 GAMLSS em análise de sobrevivência

Considere $T \sim \mathcal{D}(t; \boldsymbol{\theta}_k)$, em que $\mathcal{D}$ representa a distribuição de T , $\boldsymbol{\theta}_k$ denota o vetor de parâmetros e uma amostra de n observações independentes $t_1, \dots, t_n$. Seja c_i o tempo de censura, $y_i = \min(t_i, c_i)$ e $\Delta t_i = I(t_i \leq c_i)$, onde $\Delta t_i = 1$ se t_i é um tempo de falha, e $\Delta t_i = 0$ se for censurado à direita.

De acordo com Cordeiro, Ortega & Ramires (2015), a partir de n observações, variáveis explicativas e indicadores de censura $(t_1, \Delta t_1, x_{k1}), \dots, (t_n, \Delta t_n, x_{kn})$, o logaritmo da função de verossimilhança total sob censura não informativa para o vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)^\top$ é dada por

$$\ell(\boldsymbol{\theta}) = \sum_{i \in \mathcal{F}} \log f(t_i; \theta_i) + \sum_{i \in \mathcal{C}} \log S(t_i; \theta_i), \quad (3.15)$$

em que $S(t_i; \theta_i)$ é a função de sobrevivência relativa à distribuição $\mathcal{D}$, $\mathcal{F}$ e $\mathcal{C}$ denotam, respectivamente, os conjuntos de indivíduos cujos tempos t_i correspondem a um tempo de vida observado ou censurado e o vetor de parâmetros são definidos em 3.12 ao especificar funções de ligação apropriadas. Segundo Ramires *et al.* (2018) a maximização numérica de 3.15 pode ser realizada utilizando o pacote `gamlss` no software *R*.

A presença de fração de cura em dados de sobrevivência demanda modelos capazes de capturar, simultaneamente, os efeitos das covariáveis sob a tendência central (como a média ou mediana do tempo até o evento), a variabilidade e a probabilidade de cura. Nesse contexto, os GAMLSS (Rigby; Stasinopoulos, 2005) oferecem uma estrutura flexível, ao permitir que cada parâmetro da distribuição seja modelado por uma função de regressão distinta. As distribuições exponencial, Weibull e log-normal podem ser facilmente incorporadas, nesse contexto, ampli-

ando o poder explicativo e a capacidade de ajuste do modelo. Do ponto de vista matemático, os GAMLSS com fração de cura, baseados nessas três distribuições, são definidos como

$$\begin{aligned} g(\mu) &= X_1\beta_1 + \sum_{j=1}^{J_1} s_{1j}(x_{1j}) \\ \log(\sigma) &= X_2\beta_2 + \sum_{j=1}^{J_2} s_{2j}(x_{2j}) \\ \log\left(\frac{v}{1-v}\right) &= X_3\beta_3 + \sum_{j=1}^{J_3} s_{3j}(x_{3j}), \end{aligned} \quad (3.16)$$

em que $g(\cdot)$ indica a função logarítmica para o modelo com base na distribuição exponencial e Weibull, e identidade no caso da distribuição log-normal, X_k , $k = 1, 2, 3$, é uma matriz de delineamento, β_k é um vetor de parâmetros, $s_{kj}(\cdot)$ é uma função de suavização – por exemplo um P-spline (Eilers; Marx, 1996) – utilizada para explicar um relacionamento não-linear da covariável x_{kj} com uma das estruturas de regressão. As funções de suavização utilizadas no modelo, usualmente, não são testadas, mas sim avaliadas graficamente (Nakamura *et al.*, 2022b). Pelo fato de a distribuição exponencial só ter um parâmetro, a estrutura 3.16 irá possuir apenas os parâmetros μ e o parâmetro de fração de cura.

3.2.6 Processo de seleção do modelo

A seleção de modelos em regressão concentra-se, principalmente, na identificação de preditores relevantes a partir de um conjunto de variáveis potenciais. No caso específico de GAMLSS, o processo de seleção inclui dois aspectos fundamentais (Stasinopoulos *et al.*, 2017):

- a) a escolha da distribuição mais adequada para a variável resposta;
- b) a determinação dos preditores associados aos parâmetros da distribuição selecionada.

Seja $\mathcal{M}$ um modelo estatístico com vetor de parâmetros θ , o ajuste desse modelo pode ser avaliado por meio de sua deviência global ajustada ($GDEV$), definida como

$$GDEV = -2\ell(\theta),$$

em que $\ell(\theta)$ representa o logaritmo da função de verossimilhança ajustada, conforme descrito nas Equações 3.13, 3.14 e 3.15. Para comparar os diferentes GAMLSS ajustados, em geral,

utiliza-se o Critério de Informação Generalizado de Akaike (GAIC) (Voudouris *et al.*, 2012), que é calculado pela deviança ajustada somada a uma constante κ , proporcional aos df graus de liberdade do modelo, dado por

$$GAIC(\kappa) = GDEV + (\kappa \times df).$$

Note que, para $\kappa = 2$, o *GAIC* se reduz ao Critério de Informação de Akaike (AIC) (Akaike, 1974), enquanto, para $\kappa = \log(n)$, ele se reduz ao Critério de Informação Bayesiano (BIC) (Schwarz, 1978).

De acordo com Stasinopoulos *et al.* (2017), a escolha da distribuição adequada pode ser realizada em duas etapas: ajuste e diagnóstico. Na etapa de ajuste, os modelos ajustados com diferentes distribuições são comparados utilizando o critério *GAIC*(κ), conforme descrito anteriormente. Já a etapa de diagnóstico utiliza o *worm plot* (Buuren; Fredriks, 2001), uma ferramenta que auxilia na identificação de inadequações do modelo, seja de forma global ou em intervalos específicos de uma ou duas variáveis explicativas.

A função de ligação é escolhida, em geral, com base nos intervalos dos parâmetros e na interpretação desejada (Bastiani *et al.*, 2018). Usualmente, uma função de ligação como a identidade é utilizada quando a distribuição da resposta possui um parâmetro com valores na reta real, assim como a função logarítmica ou logito são utilizadas nos casos em que os valores dos parâmetros são positivos ou dentro do intervalo unitário, respectivamente.

Para uma distribuição específica, a seleção de termos deve ser realizada para todos os parâmetros da distribuição, não se restringindo apenas ao parâmetro de locação μ (Stasinopoulos *et al.*, 2017). Procedimentos como *forward*, *backward* e *stepwise* podem ser aplicados individualmente a cada parâmetro da distribuição. Para mais detalhes sobre esses algoritmos, recomenda-se consultar James *et al.* (2013).

De acordo com Ramires *et al.* (2021), nos GAMLSS, o processo de seleção das covariáveis para compor cada uma das estruturas de regressão do modelo mais utilizado na literatura é a denominada Estratégia A, implementada na função `stepGAICALL.A()` do pacote `gamlss`. Entretanto, conforme explicitado por Rigby & Stasinopoulos (2005), outros tipos de procedimentos podem ser adotadas, desde que bem fundamentados. Em geral, é razoável considerar que a estrutura relacionada ao parâmetro de locação seja a primeira a ser ajustada, dada sua importância para o modelo e para as interpretações. Além disso, nesta dissertação, uma vez que

serão utilizados modelos que consideram uma fração de cura – característica de suma importância para dados com esta particularidade – essa estrutura será ajustada na segunda etapa. Por fim, nos casos das distribuições Weibull e log-normal, o parâmetro relacionado à dispersão será modelado em uma em uma última etapa. Os passos utilizados são os seguintes:

- a) considerando como constantes os parâmetros relacionados à fração de cura (v) e à dispersão (σ), realiza-se um procedimento *forward* para ajustar o parâmetro de localização μ ;
- b) considerando o modelo ajustado para μ no Passo 1, e σ como uma constante, realiza-se um procedimento *forward* para ajustar o parâmetro de localização v ;
- c) considerando ambos os modelos ajustados nos Passos 1 e 2, ajusta-se um modelo para σ utilizando um último procedimento *forward*.

3.2.7 Diagnóstico do modelo

Depois que um GAMLSS é ajustado, é importante avaliar sua adequação examinando os resíduos. O uso adequado de diagnósticos é importante porque revela os pontos fortes e fracos de um modelo.

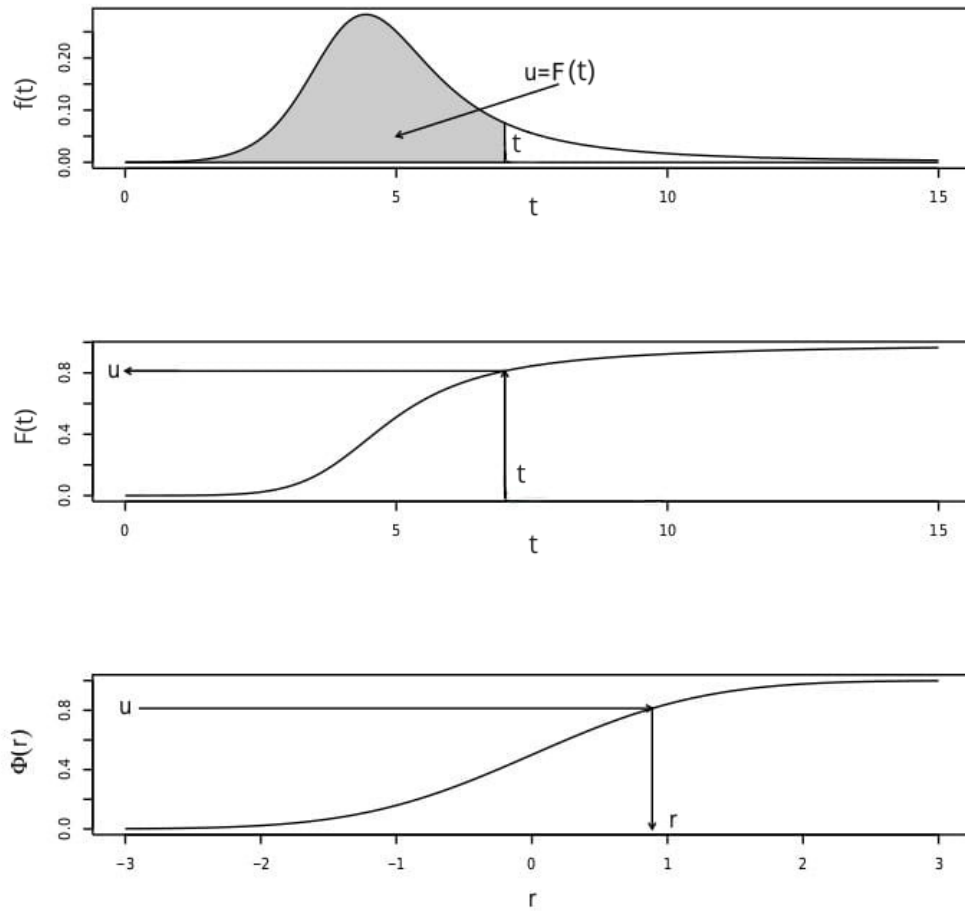
De acordo com Dunn & Smyth (1996), dado que a distribuição $f(t|\boldsymbol{\theta})$ é ajustada às observações t_i para $i = 1, \dots, n$, os resíduos quantílicos normalizados (randomizados) são dados por

$$\hat{r}_i = \Phi^{-1}(\hat{u}_i),$$

em que $\Phi^{-1}(\cdot)$ é a função inversa da função de distribuição acumulada da distribuição normal padrão. Os $\hat{u}$'s são resíduos quantílicos, definidos de forma diferente para variáveis resposta contínuas e discretas.

Se t for uma observação de uma variável resposta contínua, $u = F(t | \boldsymbol{\theta})$ e $\hat{u} = F(t | \hat{\boldsymbol{\theta}})$ representam, respectivamente, as funções de distribuição acumulada do modelo e da versão ajustada. Quando o modelo é corretamente especificado, u segue uma distribuição uniforme no intervalo $[0, 1]$, fenômeno conhecido como *probability integral transform*. Esse processo é ilustrado na Figura 3.4.

Figura 3.4 – Representação de como um resíduo quantil normalizado r é obtido para uma distribuição contínua.



Fonte: Stasinopoulos *et al.* (2017)

No painel superior, observa-se a função densidade de probabilidade (PDF) correspondente a uma observação específica t . O painel intermediário demonstra o mapeamento de t em u utilizando a função de distribuição acumulada (CDF). Por fim, no painel inferior, u é transformado em um resíduo verdadeiro r , também denominado *z-score*, conforme descrito em (Stasinopoulos *et al.*, 2017). Analogamente, $\hat{u}$ é transformado no resíduo ajustado $\hat{r}$ por meio de

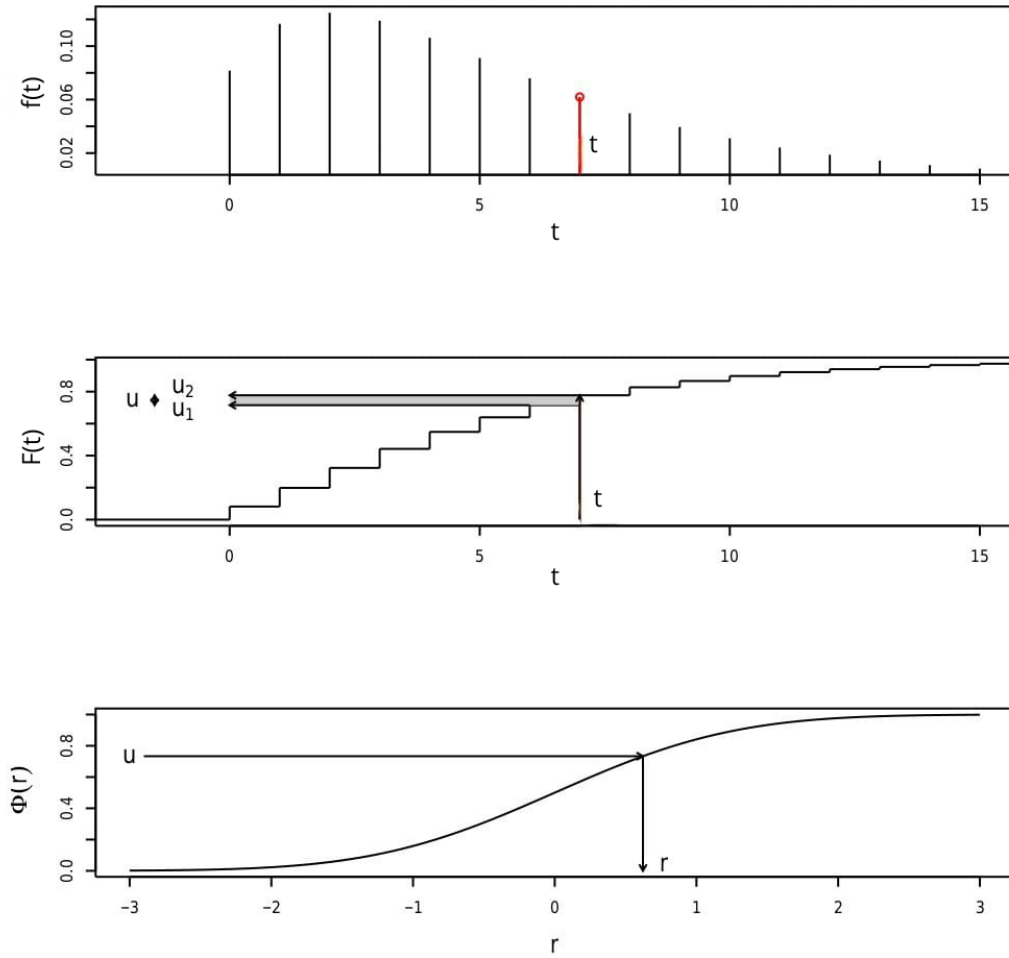
$$\hat{r} = \Phi^{-1}(\hat{u}) = \Phi^{-1} \left[F(t \mid \hat{\theta}) \right],$$

em que $\hat{r}$ também segue, aproximadamente, uma distribuição normal padrão.

Se t é uma observação de uma variável resposta discreta do tipo inteiro, $F(t \mid \theta)$ é uma função degrau com saltos nos valores inteiros $t \in R_Y$. A distribuição de $u = F(t \mid \theta)$ varia de zero a um, mas é discreta, com probabilidade positiva nos pontos $F(t \mid \theta)$. Para lidar

com essa característica, u é definido como um valor aleatório de uma distribuição uniforme no intervalo $[u_1, u_2] = [F(t-1 | \theta), F(t | \theta)]$. Da mesma forma, $\hat{u}$ é definido no intervalo $[\hat{u}_1, \hat{u}_2] = [F(t-1 | \hat{\theta}), F(t | \hat{\theta})]$ (Stasinopoulos *et al.*, 2017). Esse processo é representado na Figura 3.5.

Figura 3.5 – Representação de como um resíduo quantil randomizado normalizado r é obtido para uma distribuição discreta.



Fonte: Stasinopoulos *et al.* (2017)

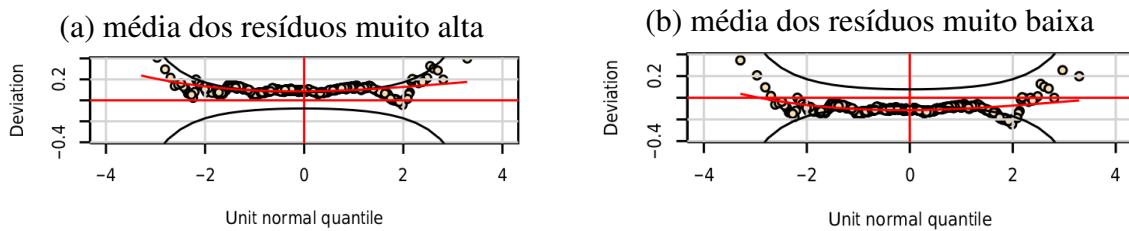
Para uma dada função de probabilidade discreta (painel superior), o valor t observado é transformado em um intervalo (u_1, u_2) , representado pela faixa sombreada no painel intermediário. Em seguida u é selecionado aleatoriamente a partir de uma distribuição uniforme no intervalo (u_1, u_2) e, posteriormente, transformado no resíduo verdadeiro (ou pontuação z), dado por $r = \Phi^{-1}(u)$ (painel inferior). A distribuição de r é exatamente normal padrão se o modelo estiver correto. De forma similar, utilizando a CDF ajustada, t é transformado em $\hat{u}$, selecionado aleatoriamente do intervalo $(\hat{u}_1, \hat{u}_2)$. Em seguida $\hat{u}$ é convertido no resíduo ajustado

$\hat{r} = \Phi^{-1}(\hat{u})$, o qual segue aproximadamente uma distribuição normal padrão (Stasinopoulos *et al.*, 2017).

O *worm plot* (Buuren; Fredriks, 2001) é uma ferramenta comum na análise de resíduos em GAMLSS, funcionando como um gráfico Q-Q normal que avalia a proximidade dos resíduos em relação à normalidade padrão.

Algumas características dos resíduos podem ser analisadas por meio do formato do *worm plot*. Quando o *worm plot* apresenta um nível abaixo da linha de origem, isso indica que os resíduos possuem uma média elevada e que o parâmetro de locação ajustado é muito baixo. Por outro lado, se o *worm plot* estiver acima da linha de origem, isso sugere que os resíduos possuem uma média reduzida e que o parâmetro de locação ajustado é muito alto (Stasinopoulos *et al.*, 2017), conforme ilustrado na Figura 3.6.

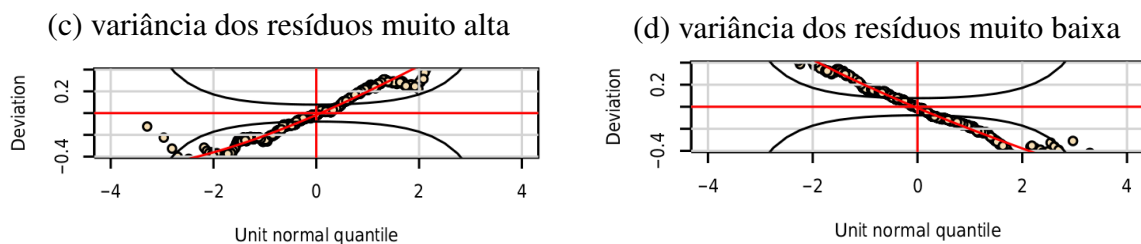
Figura 3.6 – Falhas no ajuste adequado do parâmetro de locação do modelo, indicadas pelo *worm plot*.



Fonte: Stasinopoulos *et al.* (2017)

Quando o *worm plot* exibe uma reta com inclinação positiva, isso indica que os resíduos possuem uma variância muito alta e o parâmetro de escala ajustado é muito baixo. Já quando a reta tem uma inclinação negativa, isso sugere que os resíduos apresentam uma variância reduzida e o parâmetro de escala ajustado é excessivamente alto (Stasinopoulos *et al.*, 2017), conforme mostrado na Figura 3.7.

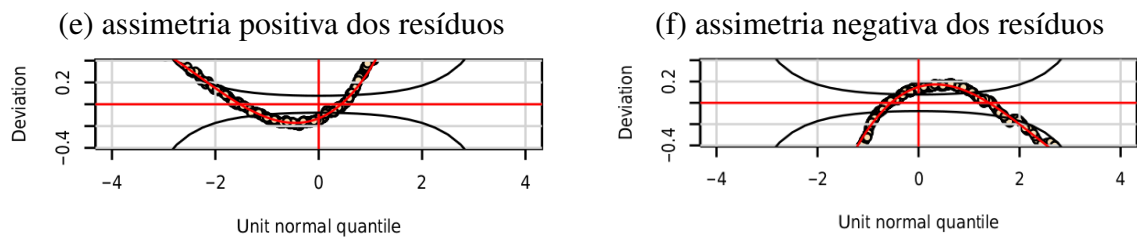
Figura 3.7 – Falhas no ajuste adequado do parâmetro de escala do modelo, indicadas pelo *worm plot*.



Fonte: Stasinopoulos *et al.* (2017)

Caso o *worm plot* apresente um formato em U, isso sugere que os resíduos exibem assimetria positiva, com o parâmetro de assimetria ajustado sendo muito baixo. Em contraste, se o *worm plot* mostrar um formato em U invertido, indica que os resíduos apresentam assimetria negativa, com o parâmetro de assimetria ajustado sendo excessivamente alto (Stasinopoulos *et al.*, 2017), conforme demonstrado na Figura 3.8.

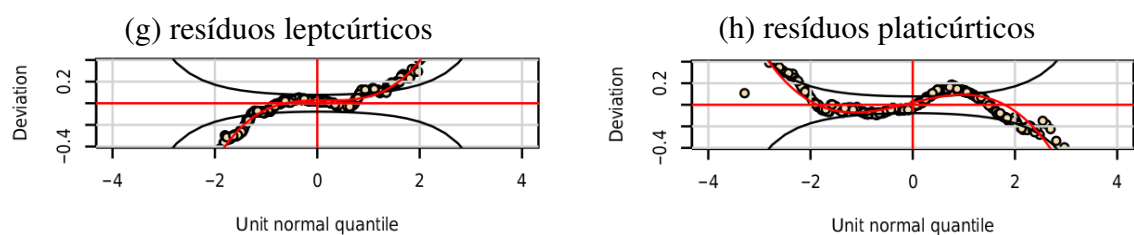
Figura 3.8 – Falhas no ajuste adequado do parâmetro de assimetria do modelo, indicadas pelo *worm plot*.



Fonte: Stasinopoulos *et al.* (2017)

Quando o *worm plot* exibe um formato em S com curvatura para baixo à esquerda, isso indica que os resíduos têm curtose leptocúrtica e as caudas da distribuição ajustada são muito leves. Caso o formato for um S com curvatura para cima à esquerda, isso sugere que os resíduos têm curtose platicúrtica e as caudas da distribuição ajustada são muito pesadas (Stasinopoulos *et al.*, 2017), conforme ilustrado na Figura 3.9.

Figura 3.9 – Falhas no ajuste adequado do parâmetro de curtose do modelo, indicadas pelo *worm plot*.



Fonte: Stasinopoulos *et al.* (2017)

Conforme Stasinopoulos *et al.* (2017), os resíduos ordenados devem estar próximos da linha horizontal pontilhada para indicar bom ajuste. Intervalos de confiança de 95% são representados por curvas elípticas, dentro das quais espera-se que 95% dos resíduos se encontrem. Uma porcentagem maior de pontos fora dessas curvas indica inadequação do modelo, e o gráfico também permite identificar *outliers*.

SEGUNDA PARTE - ARTIGO

ARTIGO 1 - Fatores prognósticos no câncer de pâncreas: uma abordagem a partir de modelos de regressão com fração de cura

Fatores prognósticos no câncer de pâncreas: uma abordagem a partir de modelos de regressão com fração de cura

Vanessa M. Silvério^a, Milena E.C. Oliveira^b, Luiz R. Nakamura^c,
Thiago G. Ramires^d, Geraldo M.C. Pereira^c

^aUniversidade Federal de Lavras, Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, Brasil.

^bCentro Universitário Tabosa de Almeida, Departamento de Psicologia, Brasil.

^cUniversidade Federal de Lavras, Departamento de Estatística, Brasil.

^dUniversidade Tecnológica Federal do Paraná, Departamento de Matemática, Brasil.

Introdução

O câncer é uma doença complexa que ocorre quando células do corpo começam a se multiplicar de forma descontrolada, podendo invadir partes adjacentes do corpo e se espalhar para outros órgãos (OPAS, 2024). Essa condição afeta milhões de pessoas no mundo e está entre as principais causas de morte prematura antes dos 70 anos, sendo um problema significativo tanto em países desenvolvidos quanto em desenvolvimento (INCA, 2023).

Dentre os diversos tipos de câncer, o de pâncreas se destaca pelo alto índice de mortalidade, apesar de sua baixa incidência. Localizado na cavidade abdominal, o pâncreas desempenha funções essenciais, como a produção de hormônios, incluindo a insulina, que regula os níveis de glicose no sangue, e a secreção de enzimas digestivas fundamentais para a absorção de nutrientes (COSTA et al., 2024).

A doença pode afetar qualquer região do pâncreas, cabeça, corpo ou cauda, e tende a ser agressiva, frequentemente infiltrando estruturas adjacentes e disseminando-se para linfonodos e órgãos distantes, como fígado e pulmões. Entre os principais fatores de risco associados à doença estão tabagismo, obesidade, histórico familiar, diabetes mellitus e pancreatite crônica (EVERETT et al., 2020).

Um dos principais desafios no enfrentamento desse tipo de câncer é a dificuldade de diagnóstico precoce. De acordo com o Instituto Nacional de Câncer - INCA (2024), nos estágios iniciais, os sintomas são inespecíficos ou inexistentes, o que leva a uma alta taxa de diagnósticos tardios, quando a doença já se encontra em estágios avançados. Os sintomas mais comuns

incluem dor abdominal, icterícia (pele e olhos amarelados), perda de peso inexplicada, náuseas, vômitos e alterações no apetite.

O estado geral do paciente no momento do diagnóstico influencia diretamente a escolha do tratamento. A cirurgia, sendo a única alternativa com potencial curativo, é indicada em poucos casos, já que a maioria dos diagnósticos ocorre em estágios avançados da doença. Quando a cirurgia não é viável, a quimioterapia e a radioterapia são utilizadas, junto com medidas de suporte para reduzir os sintomas e melhorar o bem-estar do paciente (INCA, 2024).

Embora o prognóstico geralmente seja desfavorável, alguns pacientes com câncer de pâncreas apresentam a cura, especialmente aqueles diagnosticados precocemente e submetidos a tratamentos eficazes, como a cirurgia (INCA, 2024).

O estudo do tempo de sobrevida desses pacientes pode ser realizado com base em modelos de sobrevivência, que permitem estimar a probabilidade de eventos ao longo do tempo, levando em consideração diferentes covariáveis (COLOSIMO; GIOLO, 2024). Uma das metodologias mais avançadas para esse tipo de análise é o modelo aditivo generalizado para locação, escala e forma (GAMLSS), que já foi aplicado em investigações sobre a fração de cura em diferentes tipos de câncer (RAMIRES et al., 2018; RAMIRES et al., 2019).

Sendo assim, esta pesquisa teve por objetivo identificar os fatores associados à sobrevida e à fração de cura de pacientes com câncer de pâncreas, estimando esses parâmetros com a metodologia dos modelos aditivos generalizados para locação, escala e forma - GAMLSS (RIGBY; STASINOPOULOS, 2005), por meio da distribuição Log-Normal com fração de cura.

Materiais e Métodos

Conjunto de dados

Os dados foram são provenientes da Fundação Oncocentro de São Paulo (FOSP), responsável por coordenar a base estadual do Registro Hospitalar de Câncer de São Paulo (RHC/SP) – Brasil, e referem-se a pacientes com câncer de pâncreas que foram diagnosticados no período de janeiro de 2000 a março de 2024. Após uma análise exploratória inicial, valores faltantes e variáveis com codificações desconhecidas foram identificados e removidos, resultando em um conjunto final composto por 3.683 observações. Para cada paciente, foi considerada a última informação disponível, isto é, foi obtida o tempo de sobrevivência desde o início até o final do tratamento, sendo alguns pacientes identificados como um dado censurado (33,1%), ou seja,

aqueles para os quais não foi possível determinar o momento exato do evento de interesse (morte por câncer), pois o estudo terminou antes, o paciente foi removido do acompanhamento por outros motivos ou foi curado. Além disso, apenas pacientes que não apresentaram reincidência do câncer foram considerados no estudo.

Para explicar o tempo de sobrevivência dos pacientes, foram consideradas 10 covariáveis candidatas a comporem o modelo estatístico, sendo duas classificadas como variáveis numéricas (quantitativas) e as demais como fatores (categóricas). As informações acerca de cada covariável candidata estão disponíveis na Tabela 1.

Tabela 1 – Covariáveis candidatas em estudo

Variável	Tipo	Amplitude/Categorias
Idade	Quantitativa	8 – 96 anos
Data de início do tratamento	Quantitativa	1 – 2.715 dias
Sexo	Fator	Masculino; Feminino
Estadiamento clínico	Fator	EC0: tumor está localizado, sem invasão (carcinoma in situ); EC1: Tumor restrito ao pâncreas; EC2: Tumor regional; EC3: Invasão de vasos sanguíneos principais, sem metástases; EC4: metástases distantes presentes
Tumor	Fator	T0: Lesões pré-malignas localizadas; T1: Tumor localizado, sem invasão significativa; T2: Invasão parcial de vasos importantes; T3: Invasão vascular extensa T4: Metástases distantes presentes; T5: Carcinoma in situ, sem invasão ou metástases
Linfonodos	Fator	N0: Não há metástases em linfonodos regionais; N1: Metástase em linfonodos regionais; N1A: Metástase em 1 linfonodo regional; N1B: Metástase em 2 a 3 linfonodos regionais
Metástases distantes	Fator	Não; Sim
Cirurgia	Fator	Não; Sim
Radioterapia	Fator	Não; Sim
Quimioterapia	Fator	Não; Sim

Modelagem estatística

Os modelos de mistura com fração de cura, propostos por Berkson e Gage (1952), são um dos modelos mais utilizados para lidar com situações onde pressupõe-se a existência de uma proporção de indivíduos curados ou imunes ao evento de interesse e outra proporção de indivíduos suscetíveis a ele. Nesse contexto, cada indivíduo da população é representado por um indicador N_i , $i = 1, \dots, n$, em que $N_i = 1$ indica que o indivíduo é suscetível ao evento de interesse, e $N_i = 0$ indica que é imune. Assim a população é dividida em categorias, em que indivíduos curados têm a probabilidade $0 < v < 1$, enquanto indivíduos não curados seguem uma função de sobrevivência $S(t)$ com probabilidade $(1 - v)$ (RAMIRES et al., 2019).

O modelo de mistura inclui uma função de sobrevivência que pode ser expressa como

$$S_{\text{pop}}(t_i) = v + (1 - v)S(t_i | N_i = 1), \quad (1)$$

em que $S_{\text{pop}}(t_i)$ representa a função de sobrevivência geral da população, $v = P(N_i = 0)$ indica a proporção de indivíduos curados, e $S(t_i | N_i = 1)$ descreve a função de sobrevivência a longo prazo dos indivíduos suscetíveis (RAMIRES et al., 2019).

Dentre as distribuições estatísticas mais utilizadas para se modelar problemas desta natureza, destacam-se a distribuição Weibull e log-normal (COLOSIMO; GIOLO, 2024). Se uma variável aleatória T segue uma distribuição Weibull, então sua função densidade de probabilidade (pdf) pode ser definida como

$$f(t | \mu, \sigma) = \frac{\sigma t^{\sigma-1}}{\beta^\sigma} \exp \left[- \left(\frac{t}{\beta} \right)^\sigma \right], \quad t > 0,$$

em que $\mu > 0$ é exatamente a média, $\sigma > 0$ é um parâmetro de dispersão, $\beta = \mu [\Gamma(1/\sigma + 1)]^{-1}$ e $\Gamma(z) = \int_0^\infty u^{z-1} e^{-u} du$ é a função gama (RIGBY et al., 2019). Consequentemente, sua função de sobrevivência é dada por

$$S(t | \mu, \sigma) = \exp \left\{ - \left[\frac{t \Gamma(1/\sigma + 1)}{\mu} \right]^\sigma \right\}. \quad (2)$$

Se combinarmos as equações (1) e (2) temos o modelo de sobrevivência Weibull com fração de cura (RAMIRES et al., 2018).

Por outro lado, se T segue uma distribuição log-normal, então sua fdp é escrita como

$$f(t \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi t} \sigma} \exp \left\{ -\frac{1}{2} \frac{(\log(t) - \mu)^2}{\sigma^2} \right\}, \quad t > 0,$$

em que $0 < \mu < \infty$ e $\sigma > 0$ são a média e desvio padrão de $\log(T)$. Note que, nesta parametrização, e^μ é exatamente a mediana (RIGBY et al., 2019) e, desta maneira, as interpretações podem ser realizadas de maneira mais fácil utilizando esta quantidade. Por sua vez, sua função de sobrevivência associada é dada por

$$S(t \mid \mu, \sigma) = \Phi \left(-\frac{\log(t) + \mu}{\sigma} \right), \quad (3)$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão. Ao introduzirmos a equação (3) em (1) temos o modelo de sobrevivência log-normal com fração de cura (SANTOS; ALMEIDA; OLIVEIRA, 2019).

Ambos os modelos, baseados nas distribuição Weibull e log-normal, podem ser reescritos em uma estrutura semi-paramétrica por meio dos modelos aditivos generalizados para locação, escala e forma (GAMLSS) (RIGBY; STASINOPOULOS, 2005), possibilitando assim a construção de múltiplas estruturas de regressão, capazes de identificar covariáveis que afetam significativamente a média (ou mediana) do tempo de vida do paciente, a probabilidade de cura e a variabilidade. Matematicamente, considerando que T segue uma das duas distribuições, definimos os GAMLSS com fração de cura como

$$\begin{aligned} g(\mu) &= X_1 \beta_1 + \sum_{j=1}^{J_1} s_{1j}(x_{1j}) \\ \log(\sigma) &= X_2 \beta_2 + \sum_{j=1}^{J_2} s_{2j}(x_{2j}) \\ \log\left(\frac{v}{1-v}\right) &= X_3 \beta_3 + \sum_{j=1}^{J_3} s_{3j}(x_{3j}), \end{aligned} \quad (4)$$

em que $g(\cdot)$ indica a função logarítmica para o modelo com base na distribuição Weibull e identidade no caso da distribuição log-normal, X_k , $k = 1, 2, 3$, é uma matriz de delineamento, β_k é um vetor de parâmetros, $s_{kj}(\cdot)$ é uma função de suavização – por exemplo um P-spline (EILERS; MARX, 1996) – utilizada para explicar um relacionamento não-liner da covariável

x_{kj} com uma das estruturas de regressão. As funções de suavização utilizadas no modelo, usualmente, não são testadas, mas sim avaliadas graficamente (NAKAMURA et al., 2022).

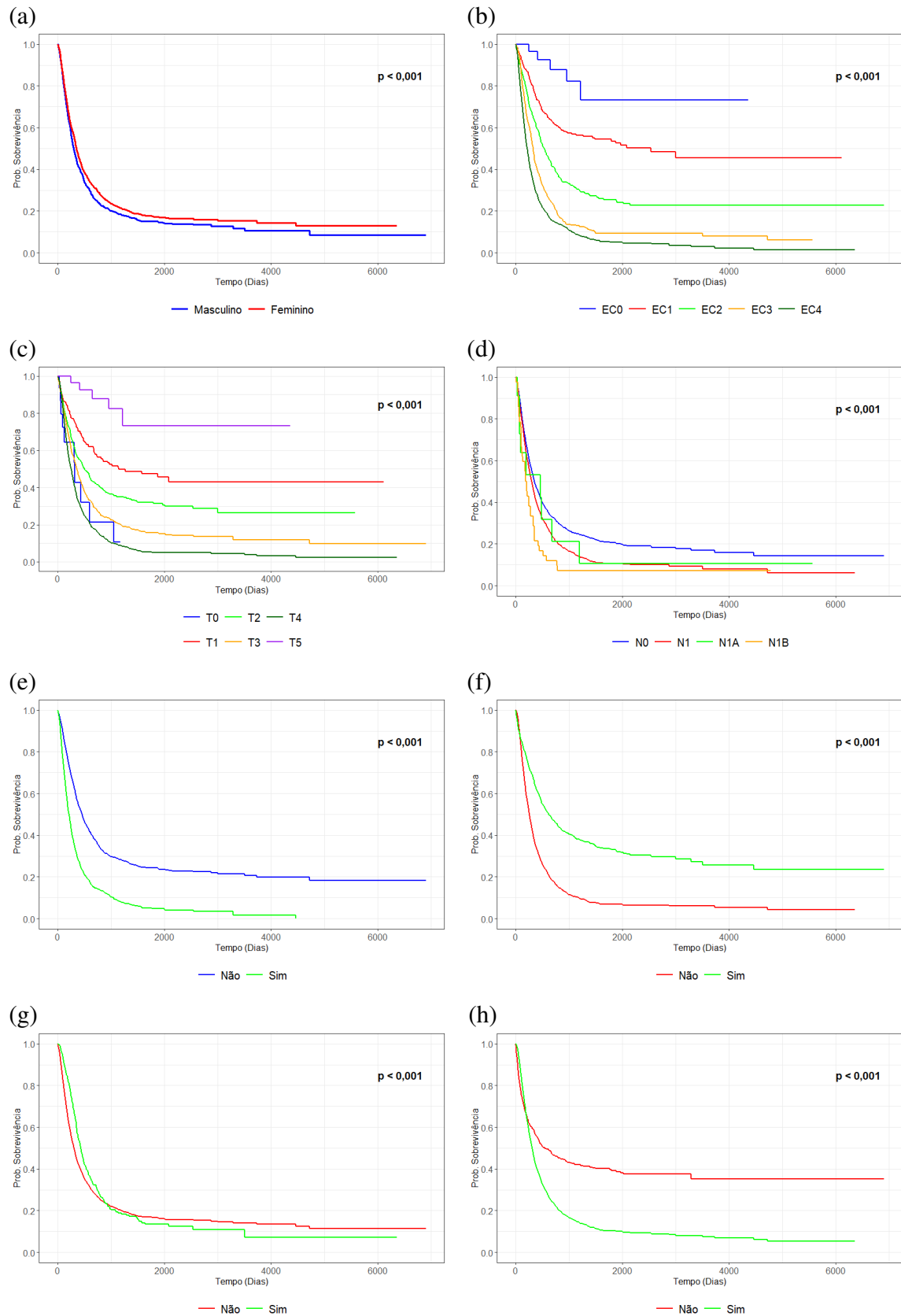
A seleção das covariáveis para compor cada uma dessas estruturas da equação (4) foi realizada da seguinte maneira: (i) um procedimento *forward* foi realizado para identificar as covariáveis que explicam o tempo de sobrevivência mediano; (ii) um novo procedimento *forward* é realizado para identificar as covariáveis que influenciam a fração de cura; (iii) um último procedimento *forward* é então realizado para identificar covariáveis na estrutura relacionada à dispersão dos tempos. Os modelos são comparados a partir dos critérios de informação de Akaike (AIC) (AKAIKE, 1974) e bayesiano (BIC) (SCHWARZ, 1978), em que menores valores implicam em melhores modelos. Finalmente, a adequação do modelo é verificada graficamente a partir de um worm plot (BUUREN; FREDRIKS, 2001) construído com base nos resíduos quantílicos normalizados (DUNN; SMYTH, 1996).

Resultados

Inicialmente, um estudo exploratório foi realizado por meio da estimação das curvas de Kaplan-Meier (KAPLAN; MEIER, 1958), nos permitindo assim, avaliar, de forma detalhada, como os fatores influenciam a probabilidade de sobrevivência ao longo do tempo, conforme ilustrado na Figura 1. Os valores-p apresentados em cada um dos painéis referem-se ao teste log-rank (MANTEL, 1966; PETO; PETO, 1972) para comparar as diferentes categorias de cada variável. Os resultados apontaram diferenças estatisticamente significativas, ao nível de 5%, entre pelo menos um par de curvas de sobrevivência para todas as variáveis categóricas analisadas, reforçando a relevância dessas características na trajetória clínica dos pacientes.

Pela Figura 1(a), observamos que o grupo com maior sobrevida é predominantemente composto por pacientes do sexo feminino, ao passo que pacientes diagnosticados em estágios iniciais da doença apresentaram maiores chances de sobrevivência em comparação com aqueles em estágios mais avançados (Figura 1(b)). Os indivíduos com tumores restritos à área primária mostraram um prognóstico mais favorável (Figura 1(c)) e, a partir da Figura 1(d), notamos que os melhores resultados relacionados ao comprometimento dos linfonodos foram observados em pacientes sem envolvimento linfático. A ausência de metástases à distância se destacou como um fator associado a uma maior sobrevida (Figura 1(e)). Pacientes que foram submetidos à cirurgia (Figura 1(f)) apresentaram uma maior sobrevida. Por outro lado, aqueles que não reali-

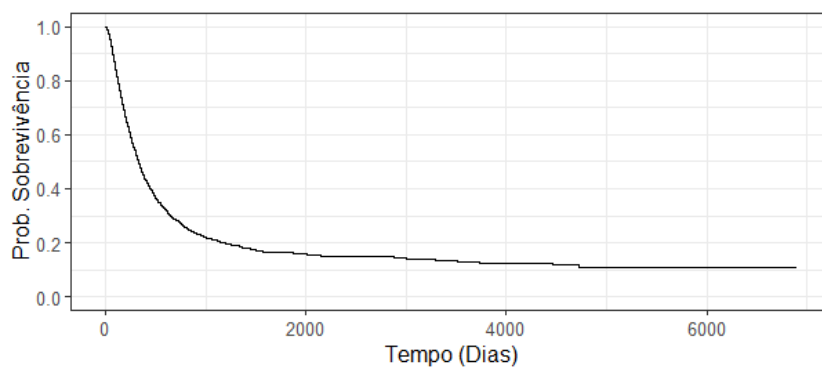
Figura 1 – Curvas de Kaplan-Meier: (a) Sexo; (b) Estadiamento clínico; (c) Tumor; (d) Linfonodos; (e) Metástase à distância; (f) Cirurgia; (g) Radioterapia; e (h) Quimioterapia.



zaram radioterapia (Figura 1(g)) e quimioterapia (Figura 1(h)) tiveram uma sobrevida superior em comparação aos que receberam tais tratamentos.

Ainda com base na Figura 1, é notório que, embora muitos pacientes venham a óbito devido ao câncer em estudo, nem todos são fatalmente afetados. Essa observação sugere a existência de uma fração de indivíduos que não está sujeita à ocorrência do evento de interesse, ou seja, uma fração de curados ou imunes ao evento (COLOSIMO; GIOLO, 2024). Essa hipótese é reforçada pela análise da função de sobrevivência estimada por meio do estimador não-paramétrico de Kaplan-Meier apresentado na Figura 2, uma vez que a curva resultante se estabiliza acima de zero.

Figura 2 – Função de sobrevivência $S(t)$



Após a análise inicial, foram ajustados os GAMLSS com fração de cura baseados nas distribuições Weibull e log-normal, utilizando a estratégia A. Os valores obtidos de AIC e BIC para os melhores modelos baseados em ambas as distribuições são apresentados na Tabela 2, onde podemos observar que o GAMLSS ajustado com base na distribuição log-normal retornou valores menores tanto para o AIC (34.764,29) quanto para o BIC (34.947,26) e, por este motivo, foi considerado como mais adequado para explicar o tempo de sobrevivência de pacientes diagnosticados com câncer de pâncreas.

Tabela 2 – Critérios de informação de Akaike (AIC) e bayesiano (BIC) obtidos para os melhores GAMLSS com fração de cura baseados nas distribuições Weibull e log-normal

Modelo	AIC	BIC
Log-normal	34.764,29	34.947,26
Weibull	35.217,42	35.396,02

Os coeficientes associados ao modelo final ajustado são apresentados na Tabela 3. Ressalta-se que os valores de e^{μ} também são apresentados para facilitar a interpretação direta do

tempo de sobrevida mediano dos pacientes, conforme explicitado na Seção Modelagem Estatística. Ainda, por meio da verificação dos resíduos do modelo construído, apresentado na Figura 3, observa-se um ligeiro formato em S dos resíduos, indicando um leve desvio em sua curtose. Uma potencial alternativa para mitigar este problema seria a utilização de distribuições mais flexíveis, isto é, onde sejamos capazes de modelar explicitamente esta característica. A utilização de modelos mais flexíveis pode comprometer a interpretabilidade do modelo, característica esta desejada na estrutura do GAMLSS (RAMIRES et al., 2021). Não obstante, observa-se que a maioria dos pontos está dentro das bandas de 95% de confiança. Sendo assim, à luz do princípio da parcimônia, podemos considerar que o ajuste obtido é adequado aos dados em estudo.

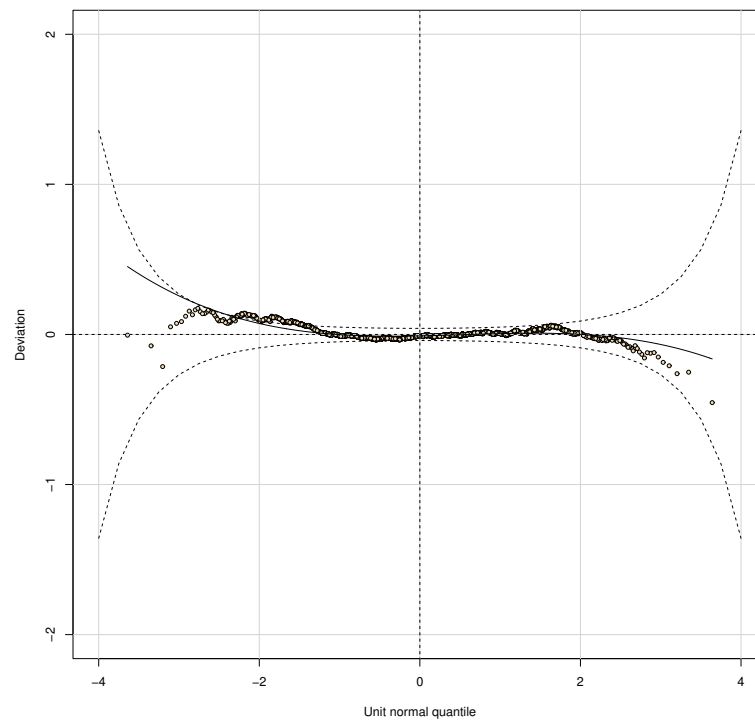
Tabela 3 – Estimativas de todas as estruturas de regressão do GAMLSS ajustado com fração de cura baseado na distribuição log-normal.

Variável	Categoria	μ	e^μ	$\log(\sigma)$	$\text{logit}(v)$
Intercepto	–	5,778	323,121	0,167	3,849
Idade	–	s_{11}^*	s_{11}^*	–	-0,048
Data do tratamento	–	s_{12}^*	s_{12}^*	–	0,002
Sexo (Referência: Masculino)	Feminino	0,087	1,090	–	–
Estadiamento clínico (Referência: EC0)	EC1	-1,005	0,366	–	-0,692
	EC2	-1,057	0,347	–	-2,041
	EC3	-1,265	0,282	–	-2,738
	EC4	-1,317	0,268	–	-3,407
Metástase à distância (Referência: Não)	Sim	-0,276	0,759	–	–
Cirurgia (Referência: Não)	Sim	0,648	1,911	0,179	–
Radioterapia (Referência: Não)	Sim	0,375	1,455	-0,257	–
Quimioterapia (Referência: Não)	Sim	0,849	2,336	-0,304	-1,447

*A função $s(\cdot)$ indica a inclusão de uma função de suavização.

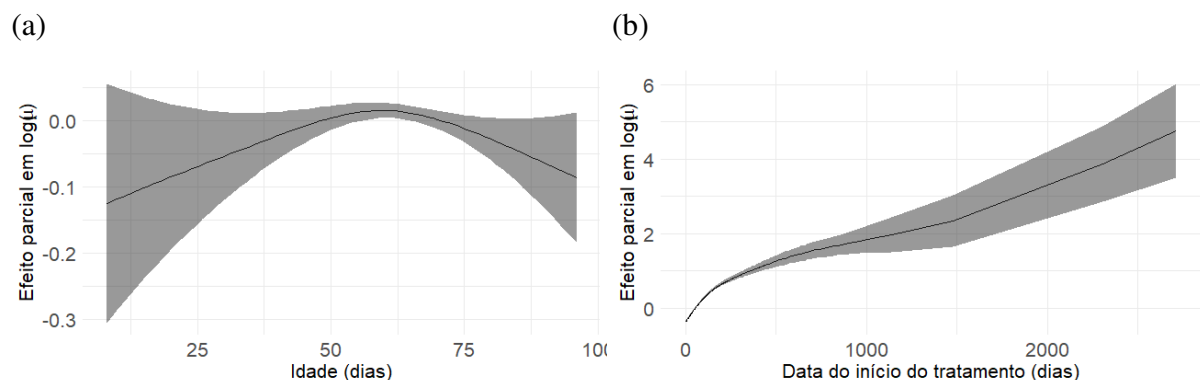
Para modelar o tempo de sobrevida mediano, definido como e^μ , foi necessário utilizar funções de suavização para capturar o relacionamento de ambas variáveis quantitativas idade e data do início do tratamento (Figura 4). Na Figura 4(a), observa-se que o tempo de sobrevivência mediano aumenta até, aproximadamente, 60 anos, e após esse limiar, esse efeito se torna negativo. A partir da Figura 4(b), constatamos que o tempo de sobrevivência mediano apresenta um aumento acentuado até, aproximadamente, 300 dias e, posteriormente, o crescimento é mais gradual.

Figura 3 – *Worm plot* obtido a partir do modelo final ajustado



Outras seis covariáveis foram selecionadas para explicar o tempo de sobrevivência mediano e^μ , todas significativas ao nível de 1%. Para a covariável sexo, é notável que mulheres apresentam um aumento de $(1,090 - 1) \times 100 = 9,0\%$ na mediana do tempo de sobrevivência em comparação aos homens. Para a covariável estadiamento clínico (ECGRUP), pacientes que possuem tumor restrito ao pâncreas (EC1) apresentam uma redução de 63,40% no tempo de sobrevivência mediano em comparação com aqueles com tumor localizado, sem invasão (carni-noma in situ – EC0). Para pacientes com tumor regional, a redução no tempo de sobrevivência

Figura 4 – Funções de suavização utilizadas para explicar o relacionamento das covariáveis (a) idade e (b) data de início do tratamento



mediano é de 65,30% em relação ao EC0. Para aqueles com invasão de vasos sanguíneos principais sem metástases e aqueles com metástases distantes, os decréscimos são de 71,80% e 73,20%, respectivamente.

No que tange à extensão da doença, pacientes com metástase à distância apresentam uma redução de 24,10% na mediana do tempo de sobrevivência, quando comparadas às pessoas que não apresentavam essa característica. Em nosso estudo, a presença dos tratamentos antineoplásicos esteve associada ao aumento no tempo de sobrevivência dos pacientes com câncer de pâncreas, no qual, pacientes que realizaram radioterapia aumentaram o tempo de sobrevivência em 45,50%, o que realizaram a cirurgia em 91,10%, e os que realizaram quimioterapia, em 133,60%.

Em relação à variabilidade dos tempos de sobrevida, identificada por meio do parâmetro σ , observamos que as variáveis relacionadas ao tratamento antineoplásico estiveram relacionadas à menor variabilidade dos tempos observados. No que tange às variáveis selecionadas na estrutura do parâmetro v , observa-se que as chances de cura aumentam nas seguintes condições: pacientes mais jovens, com tempos mais longos para início do tratamento (efeito bastante baixo desta variável), estadiamento clínico menos acentuado e, finalmente, em pacientes que não realizaram quimioterapia.

Discussão

Os resultados sugeriram que pessoas entre 8 e 60 anos apresentam um efeito positivo para a taxa de sobrevida da doença. Esse efeito se torna negativo em pessoas que apresentam mais de 60 anos (Figura 4(a)). Esses achados estão em consonância com estudos que sugerem um risco aumentado de desenvolvimento do câncer de pâncreas com a idade, especialmente, a partir dos 65 anos, sendo rara a sua ocorrência antes dos 40 anos (KLEIN, 2021). Além dos fatores relacionados à incidência, há evidências de aumento da mortalidade por câncer de pâncreas em pessoas com mais de 80 anos (KLEIN, 2021; YUAN et al., 2022; ZHANG et al., 2022; ZENG et al., 2023). Assim, considera-se que o aumento nas taxas de incidência, prevalência e mortalidade por câncer de pâncreas está relacionado com o aumento da expectativa de vida da população (KLEIN, 2021).

Um efeito positivo foi verificado em nosso estudo no que tange à data de início de tratamento no tempo de sobrevida. Contudo, Sugumar et al. (2021) realizaram um estudo de meta-análise e meta-regressão avaliando tal influência e seus resultados demonstraram uma re-

dução significativa na sobrevida global quando a terapia adjuvante foi adiada em 3 a 5 semanas após a cirurgia. Não obstante, não houve redução significativa na sobrevivência com início da terapia adjuvante em 6 a 8 semanas, ou 9 a 12 semanas, classificado como período de início tardio. [Murakami et al. \(2013\)](#), com base em um estudo com 104 prontuários clínicos de pacientes com câncer de pâncreas, concluiu uma associação da terapia adjuvante com melhor sobrevida global quando ela foi iniciada no prazo de 20 dias após a cirurgia.

Ainda, há evidências de que pacientes que receberam tratamento cinco semanas após a cirurgia tiveram sobrevida global significativamente melhor, em comparação com aqueles que iniciaram a terapia após esse ponto de corte ([KIM et al., 2017](#)). Todavia, o estudo aponta que os pacientes que não conseguiram completar seu regime de tratamento tiveram significativa piora na sobrevida. Foi demonstrado um aumento nas taxas de mortalidade entre pacientes que iniciaram o tratamento com quimioterapia adjuvante antes de 4 semanas ou após 8,4 semanas, em comparação àqueles que o iniciaram dentro do intervalo mediano entre esses períodos ([MA et al., 2019](#)). Esses dados sugerem que o início muito precoce do início do tratamento adjuvante pode colocar o paciente em risco tanto quanto o início tardio. Sendo necessário novos estudos para investigar esses dados de maneira mais robusta.

O maior nível de estadiamento foi associado ao menor tempo de sobrevida e o menor tempo de cura dos participantes avaliados. Esse dado se torna preocupante ao considerar que ainda são poucos os pacientes que recebem o diagnóstico do câncer de pâncreas no estágio I da doença ([BLACKFORD et al., 2024](#)), tornando-se necessário a construção de políticas públicas e planos de ação que visem a diminuição dessa realidade. Entre os fatores que podem ser utilizados para melhoria do diagnóstico precoce, estão: (i) melhorias nos diagnósticos por imagem, (ii) maior acesso aos serviços de saúde, otimizando as redes de cuidado no diagnóstico e tratamento e (iii) o desenvolvimento de programas de vigilância pancreática ([DBOUK et al., 2022](#)).

Pacientes com metástase apresentaram redução significativa no tempo de sobrevivência, corroborando com a literatura existente ([MIZRAHI et al., 2020](#); [BLACKFORD et al., 2024](#)). Há evidências de que o câncer de pâncreas apresenta uma biologia agressiva que corrobora para a alta mortalidade das pessoas acometidas ([GROSSBERG et al., 2020](#)). A esse respeito, estudos sugerem que 50% de pessoas com câncer de pâncreas já apresentam metástase no momento do diagnóstico e apenas 15-20% apresentam tumores ressecáveis ([WINTER et al., 2006](#)). Além disso, estima-se que a maioria dos pacientes que foram submetidos à ressecção do tumor, irá

desenvolver metástase dentro de quatro anos após a cirurgia (GROSSBERG et al., 2020; NEOPTOLEMOS et al., 2004; CONROY et al., 2018).

Em relação à variabilidade dos tempos de sobrevida, foi possível notar que as variáveis relacionadas ao tratamento antineoplásico estiveram relacionadas à menor variabilidade dos tempos observados, indicando uma possível homogeneização dos efeitos do tratamento (CHOODARI-OSKOOEI; ROYSTON; PARMAR, 2012).

No tocante aos tratamentos antineoplásicos utilizados atualmente no combate ao câncer de pâncreas, a cirurgia permanece como a principal estratégia, enquanto a quimioterapia, radioterapia e imunoterapia são opções para casos irressecáveis, embora sua eficácia ainda seja limitada devido à resistência aos medicamentos (NEOPTOLEMOS et al., 2004; STROBEL et al., 2019; HUANG et al., 2019). Em relação à cirurgia, nossos achados sugeriram maior sobrevida relacionada com o tratamento cirúrgico.

A literatura ainda apresenta discordâncias a respeito de sua associação com o sucesso terapêutico (KARUNAKARAN; BARRETO, 2021). Entre as possibilidades terapêuticas relacionadas com a cirurgia do câncer de pâncreas atualmente, sugere-se a terapia neoadjuvante, ressecções vasculares, pancreatectomia estendida, extensão da linfadenectomia e o status atual da cirurgia minimamente invasiva.

Uma revisão narrativa recente apontou que a ressecção R0, considerada potencialmente curativa, está associada a uma maior sobrevida global (KARUNAKARAN; BARRETO, 2021). Ressecções vasculares e de órgãos vizinhos devem ser realizadas apenas após terapia neoadjuvante e com confirmação, por imagens de alta qualidade, da possibilidade de atingir uma ressecção R0. A linfadenectomia estendida não mostra vantagens em relação à linfadenectomia padrão. Já as pancreatectomias distais minimamente invasivas oferecem benefícios a curto prazo, mas ainda há preocupações quanto à segurança da intervenção. Assim, recomenda-se que a cirurgia seja aplicada com rigor técnico e integrada a uma abordagem multimodal de tratamento (KARUNAKARAN; BARRETO, 2021).

O papel da terapia neoadjuvante em pacientes ressecáveis e a terapia neoadjuvante ideal de quimioterapia com ou sem radioterapia em pacientes limítrofes ressecáveis foram debatidos (WANG et al., 2024). No tocante à radioterapia, seus resultados ainda se apresentam controversos (TESSON; MORTON, 2024; EJLSMARK et al., 2023). Há algumas evidências de que a quimiorradioterapia neoadjuvante desencadeia melhor controle local de tumores em comparação com a quimioterapia sozinha (JANSSEN et al., 2022). A quimiorradioterapia também

esteve associada ao retardo da progressão local do tumor, mas sem melhora na sobrevida global (HAMMEL et al., 2016).

Por outro lado, a radioterapia neoadjuvante associada com gemcitabina melhorou a sobrevida global e a taxa de ressecção em pacientes com doença ressecável (VERSTEIJNE et al., 2022). Todavia, ainda há escassez de ensaios clínicos que avaliem diretamente a contribuição da radioterapia no ambiente perioperatório, limitando a elucidação dos dados (TESSON; MORTON, 2024).

Em casos ressecáveis ou ressecáveis limítrofes, a quimioterapia neoadjuvante pode aumentar a probabilidade de remoção cirúrgica completa das células cancerígenas. Por sua vez, em casos mais avançados, ela se apresenta como tratamento paliativo (TESSON; MORTON, 2024). Embora a quimioterapia adjuvante combinada continue sendo o padrão de tratamento para pacientes com câncer de pâncreas ressecável, a quimioterapia neoadjuvante parece melhorar a sobrevida global. Além disso, cerca de 20% dos pacientes com câncer de pâncreas não metastático irressecável podem ser submetidos à ressecção após 4 a 6 meses de quimioterapia combinada de indução com ou sem radioterapia, mesmo na ausência de uma resposta radiológica clara. Em decorrência, há um aumento na sobrevida global das pessoas que passaram por esse esquema de tratamento (SPRINGELD et al., 2023).

Contrariamente, em nosso estudo, pacientes submetidos à quimioterapia apresentaram menores chances de cura. Supõe-se que esse resultado é o reflexo de ser um tratamento utilizado em pacientes com doença mais avançada ou com metástases (TESSON; MORTON, 2024), tornando-se importante compreender que a maior sobrevida nem sempre se traduz em maiores chances de cura. Assim, sugere-se a melhor compreensão de respostas moleculares e biológicas distintas a diferentes tipos de terapias, de modo a favorecer o sequenciamento ideal de modalidades de tratamento específicas, a fim de melhorar ainda mais a sobrevida global dos pacientes com câncer de pâncreas.

Nesse ínterim, terapias mais eficazes e direcionadas têm sido investigadas, como o uso de nanopartículas, que apresentam vantagens como entrega seletiva aos tecidos tumorais, alta capacidade de carregamento de fármacos e biocompatibilidade (RAJ et al., 2021). A administração simultânea de quimioterápicos como gemcitabina e cisplatina, por meio de nanopartículas, pode exercer efeitos sinérgicos, superando barreiras da matriz tumoral e aumentando a eficácia terapêutica (TARANNUM et al., 2022). Além disso, tecnologias emergentes como a terapia fototérmica baseada em nanomateriais, especialmente o ouro, também têm se mostrado pro-

missoras ao aumentar a penetração tumoral e induzir morte celular seletiva (DUAN; LI; HE, 2023).

Considerações finais

Neste estudo, a utilização dos modelos aditivos generalizados para locação, escala e forma (GAMLSS) revelou-se adequada na modelagem dos dados relacionados ao câncer de pâncreas. O modelo proposto demonstrou um bom ajuste aos dados, evidenciando a relevância das covariáveis selecionadas para a compreensão do fenômeno em questão. A metodologia adotada permitiu uma interpretação significativa das variáveis em relação à sobrevivência e à fração de cura. Contudo, para aprimorar ainda mais os resultados, seria interessante investigar outras abordagens que possam capturar melhor a variabilidade dos dados, visando refinar a modelagem proposta.

Agradecimentos

A pesquisa foi conduzida com o apoio da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

REFERÊNCIAS

- AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.
- BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, ASA Website, v. 47, n. 259, p. 501–515, 1952. Disponível em: <<https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10501187>>.
- BLACKFORD, A. L. et al. Pancreatic cancer surveillance and survival of high-risk individuals. **JAMA Oncology**, American Medical Association, v. 10, n. 8, p. 1087, aug 2024.
- BUUREN, S. van; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. **Statistics in Medicine**, John Wiley & Sons, Ltd, England, v. 20, n. 8, p. 1259–1277, Apr 2001. ISSN 0277-6715.
- CHOODARI-OSKOOEI, B.; ROYSTON, P.; PARMAR, M. K. B. A simulation study of predictive ability measures in a survival model i: Explained variation measures. **Statistics in Medicine**, v. 31, n. 23, p. 2627–2643, oct 2012.
- COLOSIMO, E.; GIOLO, S. **Análise de Sobrevida Aplicada**. 2. ed. São Paulo: Editora Blucher, 2024. ISBN 9788521221999.
- CONROY, T. et al. Folfirinox or gemcitabine as adjuvant therapy for pancreatic cancer. **New England Journal of Medicine**, Massachusetts Medical Society, v. 379, n. 25, p. 2395–2406, dec 2018.
- COSTA, I. G. M. et al. Análise epidemiológica da neoplasia maligna de pâncreas no brasil: Internações, Óbitos e taxa de mortalidade. **Brazilian Journal of Implantology and Health Sciences**, v. 6, n. 7, p. 1007–1021, jul. 2024. Disponível em: <<https://bjih.s.emnuvens.com.br/bjih/article/view/2530>>.
- DBOUK, M. et al. The multicenter cancer of pancreas screening study: Impact on stage and survival. **Journal of Clinical Oncology**, American Society of Clinical Oncology, v. 40, n. 28, p. 3257–3266, oct 2022.
- DUAN, H.; LI, L.; HE, S. Advances and prospects in the treatment of pancreatic cancer. **International Journal of Nanomedicine**, v. 18, p. 3973–3988, jul 2023.
- DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, v. 5, n. 3, p. 236–244, 1996.
- EILERS, P. H. C.; MARX, B. D. Flexible smoothing with B-splines and penalties. **Statistical Science**, v. 11, n. 2, p. 89–121, 1996.
- EJLSMARK, M. et al. Radiotherapy for locally advanced pancreatic adenocarcinoma—a critical review of randomised trials. **Current Oncology**, v. 30, n. 7, p. 6820–6837, jul 2023.
- EVERETT et al. Pancreatic cancer: Many risk factors too poorly characterized to enable prevention. In: _____. World Cancer Report: Cancer research for cancer prevention. Lyon (FR): International Agency for Research on Cancer, 2020. Disponível em: <<https://www.ncbi.nlm.nih.gov/books/NBK606510/>>.

GROSSBERG, A. J. et al. Multidisciplinary standards of care and recent progress in pancreatic ductal adenocarcinoma. **CA: A Cancer Journal for Clinicians**, Wiley, v. 70, n. 5, p. 375–403, sep 2020.

HAMMEL, P. et al. Effect of chemoradiotherapy vs chemotherapy on survival in patients with locally advanced pancreatic cancer controlled after 4 months of gemcitabine with or without erlotinib: The lap07 randomized clinical trial. **JAMA**, v. 315, n. 17, p. 1844, may 2016.

HUANG, L. et al. Resection of pancreatic cancer in europe and usa: an international large-scale study highlighting large variations. **Gut**, BMJ Publishing Group Ltd, v. 68, n. 1, p. 130–139, jan 2019.

INCA, I. N. de C. **Estatísticas de Câncer**. 2023. <<https://www.inca.gov.br/numeros-de-cancer>>. Acesso em: 19 nov. 2024.

INCA, I. N. de C. **Estatísticas de Câncer**. 2024. <<https://www.inca.gov.br/numeros-de-cancer>>. Acesso em: 19 nov. 2024.

JANSSEN, Q. et al. Neoadjuvant radiotherapy after (m)folfirinox for borderline resectable pancreatic adenocarcinoma: A taps consortium study. **Journal of the National Comprehensive Cancer Network**, v. 20, n. 7, p. 783–791.e1, jul 2022.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, v. 53, n. 282, p. 457–481, 1958.

KARUNAKARAN, M.; BARRETO, S. G. Surgery for pancreatic cancer: current controversies and challenges. **Future Oncology**, Future Medicine Ltd, v. 17, n. 36, p. 5135–5162, 2021.

KIM, H. W. et al. Early versus delayed initiation of adjuvant treatment for pancreatic cancer. **PLoS ONE**, Public Library of Science, v. 12, n. 3, p. e0173960, mar 2017.

KLEIN, A. P. Pancreatic cancer epidemiology: understanding the role of lifestyle and inherited risk factors. **Nature Reviews Gastroenterology & Hepatology**, Nature Publishing Group, v. 18, n. 7, p. 493–502, jul 2021.

MA, S. J. et al. Association of timing of adjuvant therapy with survival in patients with resected stage i to ii pancreatic cancer. **JAMA Network Open**, American Medical Association, v. 2, n. 8, p. e199126, aug 2019.

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. **Cancer Chemother Reports**, v. 50, n. 3, p. 163–170, 1966. Disponível em: <<https://pubmed.ncbi.nlm.nih.gov/5910392/>>.

MIZRAHI, J. D. et al. Pancreatic cancer. **The Lancet**, Elsevier, v. 395, n. 10242, p. 2008–2020, jun 2020.

MURAKAMI, Y. et al. Early initiation of adjuvant chemotherapy improves survival of patients with pancreatic carcinoma after surgical resection. **Cancer Chemotherapy and Pharmacology**, Springer, v. 71, n. 2, p. 419–429, feb 2013.

NAKAMURA, L. R. et al. Cattle reference growth curves based on centile estimation: a GAMLSS approach. **Computers and Electronics in Agriculture**, v. 192, p. 106572, 2022.

- NEOPTOLEMOS, J. P. et al. A randomized trial of chemoradiotherapy and chemotherapy after resection of pancreatic cancer. **The New England Journal of Medicine**, Massachusetts Medical Society, 2004.
- OPAS. **Câncer**. 2024. Acesso em: 15 nov. 2024. Disponível em: <<https://www.paho.org/pt/topicos/cancer>>.
- PETO, R.; PETO, J. Asymptotically efficient rank invariant test procedures. **Journal of the Royal Statistical Society. Series A (General)**, v. 135, n. 2, p. 185–207, 1972.
- RAJ, S. et al. Specific targeting cancer cells with nanoparticles and drug delivery in cancer therapy. **Seminars in Cancer Biology**, v. 69, p. 166–177, feb 2021.
- RAMIRES, T. G. et al. Predicting survival function and identifying associated factors in patients with renal insufficiency in the metropolitan area of Maringá, Paraná State, Brazil. **Cadernos de Saúde Pública**, v. 34, n. 1, p. e00075517, 2018.
- RAMIRES, T. G. et al. A new semiparametric weibull cure rate model: fitting different behaviors within gamlss. **Journal of Applied Statistics**, Taylor & Francis, v. 46, n. 15, p. 2744–2760, 2019. Disponível em: <<https://doi.org/10.1080/02664763.2019.1611748>>.
- RAMIRES, T. G. et al. Validation of stepwise-based procedure in GAMLSS. **Journal of Data Science**, v. 19, n. 1, p. 96–110, 2021.
- RIGBY, R. A.; STASINOPOULOS, D. M. Generalized Additive Models for Location, Scale and Shape. **Journal of the Royal Statistical Society Series C: Applied Statistics**, v. 54, n. 3, p. 507–554, 04 2005. ISSN 0035-9254. Disponível em: <<https://doi.org/10.1111/j.1467-9876.2005.00510.x>>.
- RIGBY, R. A. et al. **Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R**. CRC Press, 2019. (Chapman & Hall/CRC The R Series). ISBN 9781000701180. Disponível em: <<https://books.google.com.br/books?id=TU-yDwAAQBAJ>>.
- SANTOS, D. F. dos; ALMEIDA, P. L. Ribeiro de; OLIVEIRA, T. A. de. Distribuição log-normal e log-normal com fração de cura para dados de sobrevivência. **Sigmae**, v. 8, n. 2, p. 323–330, Jul. 2019. Disponível em: <<https://publicacoes.unifal-mg.edu.br/revistas/index.php/sigmae/article/view/982>>.
- SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, v. 6, n. 2, p. 461–464, 1978. Disponível em: <<https://www.jstor.org/stable/2958889>>.
- SPRINGELD, C. et al. Neoadjuvant therapy for pancreatic cancer. **Nature Reviews Clinical Oncology**, v. 20, n. 5, p. 318–337, 2023.
- STROBEL, O. et al. Optimizing the outcomes of pancreatic cancer surgery. **Nature Reviews Clinical Oncology**, Nature Publishing Group, v. 16, n. 1, p. 11–26, jan 2019.
- SUGUMAR, K. et al. The importance of time-to-adjuvant treatment on survival with pancreatic cancer: A systematic review and meta-analysis. **Cancer Reports**, Wiley Online Library, v. 4, n. 5, p. e1390, oct 2021.
- TARANNUM, M. et al. Advanced nanoengineering approach for target-specific, spatiotemporal, and ratiometric delivery of gemcitabine–cisplatin combination for improved therapeutic outcome in pancreatic cancer. **Small**, v. 18, n. 2, p. 2104449, jan 2022.

TESSON, M.; MORTON, J. The preclinical gap in pancreatic cancer and radiotherapy. **Disease Models & Mechanisms**, v. 17, n. 7, p. dmm050703, jul 2024.

VERSTEIJNE, E. et al. Neoadjuvant chemoradiotherapy versus upfront surgery for resectable and borderline resectable pancreatic cancer: long-term results of the dutch randomized preopanc trial. **Journal of Clinical Oncology**, v. 40, p. 1220–1230, 2022.

WANG, J. et al. Consensus, debate, and prospective on pancreatic cancer treatments. **Journal of Hematology & Oncology**, v. 17, n. 1, p. 92, oct 2024.

WINTER, J. M. et al. 1423 pancreaticoduodenectomies for pancreatic cancer: A single-institution experience. **Journal of Gastrointestinal Surgery**, Springer, v. 10, n. 9, p. 1199–1211, nov 2006.

YUAN, C. et al. The age-dependent association of risk factors with pancreatic cancer. **Annals of Oncology**, Elsevier, v. 33, n. 7, p. 693–701, jul 2022.

ZENG, L. et al. Association of genetic risk and lifestyle with pancreatic cancer and their age dependency: a large prospective cohort study in the uk biobank. **BMC Medicine**, BioMed Central, v. 21, n. 1, p. 489, dec 2023.

ZHANG, W. et al. Nomogram predicts risk and prognostic factors for bone metastasis of pancreatic cancer: A population-based analysis. **Frontiers in Endocrinology**, Frontiers, v. 12, p. 752176, mar 2022.

TERCEIRA PARTE - CONSIDERAÇÕES FINAIS

4 CONSIDERAÇÕES FINAIS

A análise de sobrevivência com fração de cura representa uma abordagem estatística importante no contexto de doenças crônicas e de alta letalidade, como o câncer, em especial o câncer de pâncreas. Nesse cenário, a incorporação da distribuição log-normal com fração de cura no *framework* dos GAMLSS se mostrou particularmente vantajosa, ao superar restrições impostas por distribuições clássicas, que nem sempre capturam adequadamente a assimetria e a variabilidade dos tempos de sobrevivência observados.

Os resultados obtidos ao longo deste estudo evidenciam que a estrutura dos GAMLSS oferece uma grande flexibilidade na modelagem de dados com fração de cura, ao permitir a inclusão de múltiplas covariáveis e a modelagem simultânea de diferentes parâmetros da distribuição, como a média, a variabilidade e a chance de cura. O modelo log-normal proposto apresentou um bom ajuste aos dados, refletindo adequadamente a complexidade clínica associada ao câncer de pâncreas e permitindo interpretações mais detalhadas sobre os fatores associados à sobrevida dos pacientes.

Contudo, apesar do potencial dos GAMLSS, a sua aplicação no contexto da fração de cura ainda enfrenta limitações importantes, em particular, a modelagem da fração de cura dentro do *framework* GAMLSS permanece pouco desenvolvida. Além disso, não há na literatura uma análise de diagnóstico específica capaz de identificar e avaliar problemas no ajuste do modelo quando uma fração dos indivíduos é considerada curada.

Nesse sentido, estudos futuros podem contribuir de forma relevante para o avanço da área, com destaque para: (i) o desenvolvimento de extensões do GAMLSS que incorporem explicitamente a fração de cura; (ii) a proposição de novas distribuições estatísticas compatíveis com esse tipo de modelagem; (iii) a elaboração de ferramentas computacionais e pacotes específicos no software R (R Core Team, 2024); e (iv) a realização de estudos de simulação que avaliem o desempenho dos modelos sob diferentes cenários de censura e proporção de cura.

A ampliação dessas linhas de pesquisa pode ajudar a resolver as limitações atuais e incentivar o uso dos modelos GAMLSS com fração de cura em estudos clínicos e epidemiológicos mais complexos.

REFERÊNCIAS

- AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.
- BASTIANI, F. D. *et al.* Gaussian Markov random field spatial models in GAMLSS. **Journal of Applied Statistics**, v. 45, n. 1, p. 168–186, 2018.
- BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, Taylor & Francis, v. 47, n. 259, p. 501–515, 1952.
- BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. **Journal of the Royal Statistical Society: Series B**, JSTOR, v. 11, n. 1, p. 15–53, 1949.
- BUUREN, S. van; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. **Statistics in Medicine**, John Wiley & Sons, Ltd, England, v. 20, n. 8, p. 1259–1277, Apr 2001. ISSN 0277-6715.
- CASELLA, G.; BERGER, R. **Statistical Inference**. [S.l.]: Duxbury Resource Center, 2001. Textbook Binding. ISBN 0534243126.
- CASTRO, M.; CANCHO, V. G.; RODRIGUES, J. A hands-on approach for fitting long-term survival models under the gamlss framework. **Computer Methods and Programs in Biomedicine**, v. 97, p. 168–177, 2010.
- COLE, T. J.; GREEN, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. **Statistics in Medicine**, Wiley Online Library, v. 11, n. 10, p. 1305–1319, 1992.
- COLOSIMO, E.; GIOLO, S. **Análise de Sobrevivência Aplicada**. 2. ed. São Paulo: Editora Blucher, 2024. ISBN 9788521221999.
- CORDEIRO, G. M.; ORTEGA, E. M. M.; RAMIRES, T. G. A new generalized weibull family of distributions: mathematical properties and applications. **Journal of Statistical Distributions and Applications**, v. 2, n. 1, p. 13, Nov 2015. ISSN 2195-5832. Disponível em: <https://doi.org/10.1186/s40488-015-0036-6>.
- DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, v. 5, n. 3, p. 236–244, 1996. ISSN 10618600.
- EILERS, P. H. C.; MARX, B. D. Flexible smoothing with B-splines and penalties. **Statistical Science**, v. 11, n. 2, p. 89–121, 1996.
- HARRELL, F. **Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis**. [S.l.]: Springer International Publishing, 2015. (Springer Series in Statistics). ISBN 9783319194240.
- HASTIE, T.; TIBSHIRANI, R. **Generalized additive models**. [S.l.]: Chapman and Hall/CRC, 1990.

HELLER, G. Z.; ROBLEDOS, K. P.; MARSCHNER, I. C. Distributional regression in clinical trials: treatment effects on parameters other than the mean. **BMC Medical Research Methodology**, BioMed Central, v. 22, n. 1, p. 56, Feb 2022. ISSN 1471-2288. Disponível em: <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01534-8>.

INCA, I. N. de C. **Estatísticas de Câncer**. 2023. <https://www.inca.gov.br/numeros-de-cancer>. Acesso em: 19 nov. 2024.

JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in R**. Springer, 2013. Disponível em: <https://faculty.marshall.usc.edu/gareth-james/ISL/>.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, [American Statistical Association, Taylor and Francis, Ltd.], v. 53, n. 282, p. 457–481, 1958. ISSN 01621459, 1537274X. Disponível em: <http://www.jstor.org/stable/2281868>.

MALLER, R. A.; ZHOU, X. **Survival analysis with long-term survivors**. Chichester, Inglaterra: Wiley, 1996. 278 p. (Wiley Series in Probability and Statistics).

MOORE, D. **Applied Survival Analysis Using R**. [S.l.]: Springer International Publishing, 2016. (Use R!). ISBN 9783319312453.

NAKAMURA, L. R. *et al.* Using the box-cox family of distributions to model censored data: a distributional regression approach. **Brazilian Journal of Biometrics**, v. 40, n. 4, p. 407–414, Dec. 2022. Disponível em: <https://biometria.ufla.br/index.php/BBJ/article/view/625>.

NAKAMURA, L. R. *et al.* Cattle reference growth curves based on centile estimation: a GAMLSS approach. **Computers and Electronics in Agriculture**, v. 192, p. 106572, 2022.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society. Series A (General)**, [Royal Statistical Society, Oxford University Press], v. 135, n. 3, p. 370–384, 1972. ISSN 00359238, 23972327. Disponível em: <http://www.jstor.org/stable/2344614>.

PAL, S. *et al.* A support vector machine-based cure rate model for interval censored data. **Statistical Methods in Medical Research**, SAGE Publications, v. 32, n. 12, p. 2405–2422, dez. 2023. ISSN 1477-0334. Disponível em: <https://doi.org/10.1177/09622802231210917>.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2024. Disponível em: <https://www.R-project.org/>.

RAMIRES, T. *et al.* Validation of stepwise-based procedure in gamlss. **Journal of Data Science**, p. 96–110, 01 2021.

RAMIRES, T. G. *et al.* Predicting survival function and identifying associated factors in patients with renal insufficiency in the metropolitan area of maringá, paraná state, brazil. **Cadernos de Saúde Pública**, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, v. 34, n. 1, p. e00075517, 2018. ISSN 0102-311X. Disponível em: <https://doi.org/10.1590/0102-311X00075517>.

RAMIRES, T. G. *et al.* A new semiparametric weibull cure rate model: fitting different behaviors within gamlss. **Journal of Applied Statistics**, v. 46, p. 2744–2760, 2019.

RIGBY, R. A.; STASINOPOULOS, D. A semi-parametric additive model for variance heterogeneity. **Statistics and Computing**, Springer, v. 6, p. 57–65, 1996.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized Additive Models for Location, Scale and Shape. **Journal of the Royal Statistical Society Series C: Applied Statistics**, v. 54, n. 3, p. 507–554, 04 2005. ISSN 0035-9254. Disponível em: <https://doi.org/10.1111/j.1467-9876.2005.00510.x>.

RIGBY, R. A. *et al.* **Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R**. [S.l.]: CRC Press, 2019.

ROCHA, C.; PAPOILA, A. L. Análise de sobrevivência. In: **Congresso Anual da Sociedade Portuguesa de Estatística, XVII**. Lisboa: INE, 2009. ISBN 978-972-8890-22-3. Disponível em: <https://www.ine.pt/xurl/pub/148664525>.

SAÚDE, O. P.-A. da. **Câncer**. 2024. Acesso em: 15 nov. 2024. Disponível em: <https://www.paho.org/pt/topicos/cancer>.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. ISSN 00905364, 21688966. Disponível em: <http://www.jstor.org/stable/2958889>.

STASINOPOULOS, D.; RIGBY, R. Generalized additive models for location scale and shape (gamlss) in r. **Journal of Statistical Software**, v. 23, 12 2007.

STASINOPOULOS, D. M. *et al.* **Flexible regression and smoothing: Using GAMLSS in R**. [S.l.]: Chapman and Hall/CRC, 2017. ISBN 9781138197909.

STASINOPOULOS, M. D.; RIGBY, R. A.; BASTIANI, F. D. Gamlss: A distributional regression approach. **Statistical Modelling**, v. 18, n. 3-4, p. 248–273, 2018. Disponível em: <https://doi.org/10.1177/1471082X18759144>.

VOUDOURIS, V. *et al.* Modelling skewness and kurtosis with the BCPE density in GAMLSS. **Journal of Applied Statistics**, v. 39, n. 6, p. 1279–1293, 2012.

WEIBULL, W. **A Statistical Theory of the Strength of Materials**. Stockholm: Generalstabens Litografiska Anstalts Förlag, 1939.

APÊNDICES

APÊNDICE A – CÓDIGO PARA A IMPLEMENTAÇÃO DA DISTRIBUIÇÃO LOG-NORMAL COM FRAÇÃO DE CURA.

Este apêndice apresenta os códigos em R utilizados neste trabalho para implementação da distribuição log-normal com fração de cura no GAMLSS, os quais também estão disponíveis no repositório https://github.com/VanessaSilverio/GAMLSS_e_Log_Normal_cr.git

```
LOGNOcr <- function (mu.link="identity", sigma.link="log", nu.
link="logit") {
  mstats <- checklink("mu.link", "Log_Normal_cr", substitute(
    mu.link), c("inverse", "log", "identity", "own"))
  dstats <- checklink("sigma.link", "Log_Normal_cr",
    substitute(sigma.link), c("inverse", "log", "identity",
    "own"))
  vstats <- checklink("nu.link", "Log_Normal_cr", substitute(
    nu.link), c("logit", "probit", "cloglog", "cauchit", "
    log", "own"))

  structure(
    list(family = c("LOGNOcr", "Log_Normal_cure_rate"),
      parameters = list(mu=TRUE, sigma=TRUE, nu=TRUE),
      nopar = 3,
      type = "Continuous",

      nu.link = as.character(substitute(nu.link)),
      nu.linkfun = vstats$linkfun,
      nu.linkinv = vstats$linkinv,
      nu.dr = vstats$mu.eta,

      mu.link = as.character(substitute(mu.link)),
      sigma.link = as.character(substitute(sigma.link)),
      mu.linkfun = mstats$linkfun,
      sigma.linkfun = dstats$linkfun,
      mu.linkinv = mstats$linkinv,
      sigma.linkinv = dstats$linkinv,
      mu.dr = mstats$mu.eta,
      sigma.dr = dstats$mu.eta,

      dldm = function(y,mu,sigma) {dldm <- (log(y)-mu)/sigma^2;
        dldm},
      d2ldm2 = function(sigma) -1/sigma^2,
```

```

d1dd = function(y,mu,sigma) {d1dd <- (1/(sigma^3))*((log
  (y)-mu)^2-sigma^2); d1dd},
d2l1dd2 = function(sigma) -2/sigma^2,
d2l1dmdd = function(y) rep(0,length(y)),

d1dv = function(y,nu) -1/(1-nu),
d2l1dv2 = function(y,nu) -1/(1-nu)^2,
d2l1dmddv = function(y,mu,sigma,nu){
  nd = gamlss:::numeric.deriv(dLOGNOcr(y, mu, sigma, nu,
    log = TRUE), "mu", delta = 1e-04)
  d1dm = as.vector(attr(nd, "gradient"))
  d2l1dmddv = -d1dm * (-1/(1-nu))
  d2l1dmddv},
d2l1dddv = function(y,mu,sigma,nu){
  nd = gamlss:::numeric.deriv(dLOGNOcr(y, mu, sigma, nu,
    log = TRUE), "sigma", delta = 1e-04)
  d1dm = as.vector(attr(nd, "gradient"))
  d2l1dmddv = -d1dm * (-1/(1-nu))
  d2l1dmddv},

G.dev.incr = function(y,mu,sigma,nu,...) -2*dLOGNOcr(x=y
  , mu=mu, sigma = sigma, nu=nu,log = TRUE), #
  rqres = expression( rqres(pfuns="pLOGNOcr",
    type="Continuous", y=y, mu=mu, sigma=
      sigma,nu=nu)),
  mu.initial = expression({ mu <- (log(y)+mean(log(
    y)))/2 } ),
  sigma.initial = expression({sigma <- rep(sd(log(y)),
    length(y)) } ),
  mu.valid = function(mu) all(mu > 0),
  sigma.valid = function(sigma) all(sigma > 0),
  nu.initial = expression(nu <-rep(0.5, length(y))),
  nu.valid = function(nu) all(nu >= 0.001) && all(nu < .99)
  ,

y.valid = function(y) all(y > 0)
),
  class = c("gamlss.family","family"))
}
#-----

dLOGNOcr <- function(x, mu=0, sigma=1,nu=0, log = FALSE)
{
  if (any(sigma <= 0) ) stop(paste("sigma_must_be_
    greater_than_0_", "\n", ""))
  if (any(x < 0) ) stop(paste("x_must_be_>=0", "\n", "
    "))

```

```

        fy1 <- (1-nu)*dlnorm(x=x, meanlog = mu, sdlog = sigma
        )
        if(log==FALSE) fy<-fy1 else fy<-log(fy1)
        fy}

#-----
pLOGNOcr <- function(q, mu=0, sigma=1, nu=0, lower.tail = TRUE,
log.p = FALSE)
{
  #      if (any(mu <= 0) ) stop(paste("mu must be greater
  than 0 ", "\n", ""))
  if (any(sigma <= 0) ) stop(paste("sigma_must_be_
  greater_than_0_", "\n", ""))
  if (any(q < 0) ) stop(paste("y_must_be_>=0", "\n", "
  "))
  cdf1 <- (1-nu)*plnorm(q=q, meanlog = mu, sdlog =
  sigma)
  if(lower.tail==TRUE) cdf<-cdf1 else cdf<- 1-cdf1
  if(log.p==FALSE) cdf<- cdf else cdf<- log(cdf)
  cdf
}

#-----
qLOGNOcr <- function(p, mu=0, sigma=1, nu=0, lower.tail = TRUE,
log.p = FALSE )
{
  #if (any(mu <= 0) ) stop(paste("mu must be greater than 0
  ", "\n", ""))
  if (any(sigma <= 0) ) stop(paste("sigma_must_be_
  greater_than_0_", "\n", ""))
  if (any(p < 0) | any(p > 1)) stop(paste("p_must_be_
  between_0_and_1", "\n", ""))
  if (any(p < nu)) stop(paste("p_must_be_less_than_nu"
  , "\n", ""))

  q <-qlnorm(p=p/(1-nu), meanlog = mu, sdlog = sigma,
  lower.tail = lower.tail, log.p = log.p)
  q
}

#-----
rLOGNOcr <- function(n, mu=0, sigma=1, nu=0)
{
  if (any(sigma <= 0) ) stop(paste("sigma_must_be_
  greater_than_0_", "\n", ""))
  if (any(n <= 0)) stop(paste("n_must_be_a_positive_
  integer", "\n", ""))

  n2=round(nu*n,0) #cure

```

```

n1=n-n2 #failure

unif1<- runif(n = n1,0,1) # failure
r1 <- rlnorm(unif1,mu,sigma) #failure

dp<-sd(r1);
r2 <- runif(n2,max(r1)+0.5*dp,max(r1)+2*dp) #cure

Response<-c(r1,r2)

delta1<-rep(1,n1); #failure
delta2<-rep(0,n2) #censure
Delta<-c(delta1,delta2)
print("Variables_Response_and_Delta(1=failure,_0=censored)_has
      _been_successfully_generated")
}
#-----

```

<https://bibliotecauniversitaria.ufla.br/servicos-biblioteca/manual-de-normalizacao>

APÊNDICE B – CÓDIGO PARA AJUSTE DO GAMLSS UTILIZANDO A DISTRIBUIÇÃO LOG-NORMAL COM FRAÇÃO DE CURA.

Este apêndice apresenta os códigos em R utilizados para o ajuste do modelo GAMLSS utilizando a distribuição log-normal com fração de cura, também disponibilizados no repositório https://github.com/VanessaSilverio/GAMLSS_e_Log_Normal_cr.git

```

m1 <-gamlss(Surv(diassobrev, ULTINFO) ~ 1,
            family= cens('LOGNO2cr'),data=dados1)
nC <- detectCores(); nC
m1 <- stepGAIC(m1, scope=expl, parallel = 'multicore', ncpus =
  nC, direction = 'forward')
m1 <- stepGAIC(m1, scope=expl, parallel = 'multicore', ncpus =
  nC, direction = 'forward', what = 'nu')
~QUIMIO + ECGRUP + pb(IDADE) + CIRURGIA + RADIO + pb(datatrat)
m1 <- stepGAIC(m1, scope=expl, parallel = 'multicore', ncpus =
  nC, direction = 'forward', what = 'sigma')
~QUIMIO + RADIO + CIRURGIA + pb(IDADE)

mLOGNO <- gamlss(Surv(diassobrev, ULTINFO) ~ ECGRUP + pb(
  datatrat) + CIRURGIA +
  QUIMIO + RADIO + pb(IDADE) + SEXO + M, #
  removi T

```

```
sigma.formula = ~QUIMIO + RADIO + CIRURGIA, #  
             removi idade  
nu.formula = ~QUIMIO + ECGRUP + IDADE +  
             datatrat, # removi radio e cirurgia  
family = cens("LOGNOcr"),  
data = dados1)
```