Universidade Federal da Bahia - UFBA
Instituto de Matemática e Estatística - IME
Programa de Pós-Graduação em Matemática - PGMAT
Dissertação de Mestrado

A família estendida logística generalizada dependente do tempo - GTDEL

Ângela Lima da Silva

Salvador - Bahia
Julho de 2025

A família estendida logística generalizada dependente do tempo - GTDEL

Ângela Lima da Silva

Dissertação de Mestrado apresentada ao Colegiado da Pós-Graduação em Matemática da Universidade Federal da Bahia (UFBa), como parte dos requisitos para obtenção do título de Mestre em Matemática. Área de Concentração: Estatística.

Orientador: Prof. Dr. Jalmar Manuel Farfan Carrasco.

Salvador - Bahia
Julho de 2025

Ficha catalográfica elaborada pela Biblioteca Universitária de Ciências e Tecnologias Prof. Omar Catunda, SIBI – UFBA.

S586 Silva, Ângela Lima da

A família estendida logística generalizada dependente do tempo - GTDEL/ Ângela Lima da Silva. – Salvador, 2025.

104 f.

Orientador: Prof. Dr. Jalmar Manuel Farfan Carrasco

Dissertação (Mestrado) – Universidade Federal da Bahia. Instituto de Matemática e Estatística, 2025.

1. GTDL. 2. GTDEL. 3. Monte Carlo. 4. pbc. I. Carrasco, Jalmar Manuel Farfan. II. Universidade Federal da Bahia. III. Título.

CDU: 519.2(043.3)

A família estendida logística generalizada dependente do tempo - GTDEL

Angela Lima da Silva

Dissertação apresentada ao Colegiado do Curso de Pós-graduação em Matemática da Universidade Federal da Bahia, como requisito parcial para obtenção do Título de Mestre em Matemática.

Banca examinadora

Documento assinado digitalmente
gov.br
JALMAR MANUEL FARFAN CARRASCO
Data: 01/09/2025 10:28:58-0300
Verifique em https://validar.iti.gov.br

Profº Drº Jalmar Manuel Farfan Carrasco (orientador - UFBA)

Documento assinado digitalmente
gov.br
GIOVANA OLIVEIRA SILVA
Data: 01/09/2025 14:23:36-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Giovana Oliveira Silva (UFBA)

Documento assinado digitalmente
gov.br
ANA CARLA PERCONTINI DA PAIXAO
Data: 01/09/2025 11:00:39-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Ana Carla Percontini da Paixão (UEFS)

Agradeço a Deus.

Agradecimentos

A Deus, fonte de força, sabedoria e perseverança, agradeço por me guiar durante toda essa jornada, iluminando meu caminho nos momentos de dificuldade e me concedendo a oportunidade de seguir em frente com fé e determinação.

À Universidade Federal da Bahia (UFBA) e ao Programa de Pós-Graduação em Matemática, sou grata pelo ambiente acadêmico enriquecedor e pelas oportunidades que me foram oferecidas ao longo do mestrado. Meu reconhecimento também à Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) pelo apoio financeiro, que foi essencial para a realização deste trabalho.

Aos professores do Instituto de Matemática, que compartilharam seus conhecimentos e me ajudaram a crescer academicamente, e aos colegas e funcionários, cujo convívio e colaboração tornaram essa trajetória mais leve e produtiva, expresso minha sincera gratidão.

Um agradecimento especial ao meu orientador, Professor Dr.Jalmar Carrasco, por sua paciência, incentivo e dedicação. Além de excelente orientador, é uma pessoa maravilhosa, amigo e possuidor de imenso saber. Sua orientação cuidadosa, sua disposição para discutir ideias e seu apoio incondicional foram fundamentais para a conclusão desta dissertação.

De maneira igualmente especial, agradeço às professoras que compuseram a banca examinadora, Profa. Dra. Giovana Oliveira Silva (UFBA) e Profa. Dra. Ana Carla Percontine da Paixão, da Universidade Estadual de Feira de Santana (UEFS), pela leitura atenta, pelas contribuições valiosas e pela generosidade em compartilhar seus conhecimentos. Suas observações e sugestões foram fundamentais para o aprimoramento deste trabalho.

Por fim, um agradecimento aos meus familiares, que sempre estiveram ao meu lado oferecendo apoio emocional, compreensão e incentivo nos momentos mais desafiadores. Sem vocês, essa conquista não teria sido possível.

Muito Obrigada!
Ângela Lima

Busco no novo sem medo do passado.
Marcelo D2

Sumário

Resumo

Estudos demostram que certos conjuntos de dados de sobrevivência não são adequadamente representados pelo modelo de riscos proporcionais de Cox (1972), evidenciando a necessidade de abordagens alternativas que acomodem a não proporcionalidade dos riscos. Essa limitação tem motivado o desenvolvimento de modelos que ampliam as possibilidades analíticas na análise de sobrevivência.

Nesse contexto, destaca-se a família logística generalizada dependente do tempo (GTDL), alternativa promissora ao modelo de Cox por permitir a modelagem de situações em que a suposição de riscos proporcionais é violada. Contudo, uma limitação relevante do modelo GTDL é a sua incapacidade de acomodar a forma em banheira da função de risco, padrão frequentemente observado em dados empíricos.

Diante disso, este trabalho propõe uma extensão do modelo GTDL, denominada família estendida logística generalizada dependente do tempo (GTDEL), capaz de representar uma gama mais ampla de comportamentos da função de risco, incluindo a forma em banheira. Apresenta-se a formulação matemática da nova distribuição, suas principais propriedades teóricas e a construção do modelo de regressão associado.

Na literatura, o modelo GTDL é geralmente apresentado para o caso em que o parâmetro α é positivo. Neste estudo, enfatiza-se a importância de investigar o comportamento do modelo na situação em que $\alpha < 0$, introduzindo assim as versões denominadas modelos GTDL e GTDEL defeituosos.

Os estimadores de máxima verossimilhança são obtidos e seu desempenho assintótico é avaliado por meio de estudos de simulação Monte Carlo, sob diferentes cenários experimentais. Por fim, aplica-se a proposta a quatro conjuntos de dados reais, com destaque para o conjunto `pbc`, oriundo de um estudo clínico sobre pacientes com cirrose biliar primária.

Palavras-chave: GTDL, GTDEL, Monte Carlo, `pbc`.

Abstract

Studies have shown that certain survival datasets are not adequately represented by the proportional hazards model proposed by Cox (1972), highlighting the need for alternative approaches that accommodate non-proportional hazards. This limitation has driven the development of models that expand the analytical possibilities in survival analysis. In this context, the Generalized Time-Dependent Logistic (GTDL) family stands out as a promising alternative to the Cox model, as it allows modeling scenarios in which the proportional hazards assumption is violated. However, a relevant limitation of the GTDL family is its inability to accommodate the bathtub-shaped hazard function, a pattern frequently observed in empirical data. To address this issue, this work proposes an extension of the GTDL family, called the Extended Generalized Time-Dependent Logistic (GTDEL) family, which is capable of representing a broader range of hazard function shapes, including the bathtub form. The mathematical formulation of the new distribution is presented, along with its main theoretical properties and the construction of the associated regression model. In the literature, the GTDL model is usually considered in the case where the parameter α is positive. In this study, we emphasize the importance of investigating the model's behavior when $\alpha < 0$, thereby introducing the so-called defective versions of the GTDL and GTDEL models. Maximum likelihood estimators are derived, and their asymptotic performance is assessed through Monte Carlo simulation studies under different experimental scenarios. Finally, the proposed methodology is applied to four real datasets, among which the `pbc` dataset stands out, originating from a clinical study involving patients with primary biliary cirrhosis.

Keywords: GTDL, GTDEL, Monte Carlo, `pbc`..

LISTA DE FIGURAS

LISTA DE TABELAS

Capítulo 1

Introdução

Ao trabalhar com modelos probabilísticos, é necessário que se internalize sua condição de aproximação da realidade, ou seja, não existem modelos probabilísticos exatos, pois sempre haverá algum grau de perda de informação devido à presença inerente de componentes aleatórios, como erros de medição, flutuações naturais ou variabilidade intrínseca aos fenômenos estudados. Essa condição é intrínseca à natureza dos modelos probabilísticos, que buscam capturar padrões e tendências em meio à heterogeneidade dos dados.

Esse processo de compreensão e aprimoramento dos modelos probabilísticos está profundamente enraizado na evolução da teoria das probabilidades, que estabeleceu as bases para os métodos estatísticos modernos. Ao fornecer uma estrutura matemática sólida, a teoria das probabilidades permitiu modelar incertezas e variabilidades observadas em fenômenos naturais e sociais, abrindo caminho para o desenvolvimento de técnicas cada vez mais sofisticadas.

Nas últimas décadas, o avanço das teorias estatísticas e computacionais ampliou significativamente as possibilidades de investigação, permitindo a criação de novos modelos que aprimoram a capacidade de representar fenômenos complexos. Nesse contexto, as distribuições de probabilidade desempenham um papel central, sendo uma das principais áreas de estudo na Estatística. Contudo, os modelos existentes nem sempre se ajustam adequadamente aos conjuntos de dados analisados, o que incentiva o desenvolvimento de novas abordagens. Essas novas formulações são frequentemente empregadas para proporcionar um melhor ajuste aos dados, especialmente em situações em que os modelos clássicos falham em capturar a complexidade dos fenômenos observados.

Um exemplo significativo desse avanço é a modelagem de dados clínicos e industriais, que é realizada em Análise de Sobrevivência. Essa é uma área fundamental da estatística que tem como objetivo estudar o tempo até a ocorrência de um determinado evento, como a morte de um paciente, a falha de um componente mecânico ou a progressão de uma doença. Modelos estatísticos são empregados para descrever a distribuição desses tempos e identificar fatores que influenciam a duração até o evento.

A introdução do modelo de riscos proporcionais, de Cox (1972), empregado em Análise de Sobrevivência, revolucionou a maneira como estudamos fenômenos como o tempo de duração de componentes, a recorrência de doenças e o tempo até a falha de equipamentos. Esse modelo pressupõe que a razão de riscos entre diferentes grupos se mantém constante ao longo do tempo, o que contribui para sua ampla adoção. No entanto, como destacado por Struthers e Kalbfleisch (1986), quando essa suposição é violada nos dados, as estimativas dos coeficientes podem se tornar enviesadas. Por isso, é fundamental aplicar testes que verifiquem a validade da suposição de proporcionalidade antes de ajustar o modelo.

Para lidar com situações em que a suposição de proporcionalidade dos riscos não se sustenta, foram desenvolvidos modelos alternativos, como o modelo estratificado de taxas de falha proporcionais (Colosimo, 1997), os modelos de riscos multiplicativos dinâmicos (Therneau e Grambsch, 2000), o modelo de riscos aditivos para dados agrupados (Aranda-Ordaz e J, 1983; Tibshirani,

1983), e o modelo parcialmente paramétrico (McKeague e Sasieni, 1994). Essas abordagens oferecem maior flexibilidade para representar a dinâmica dos dados quando a suposição de riscos proporcionais é violada.

Dentre essas alternativas, destaca-se o Modelo Logístico Generalizado Dependente do Tempo (GTDL), proposto por Mackenzie (1996), que incorpora uma função de risco com dependência explícita do tempo. Ao flexibilizar a estrutura da função de risco por meio de uma parametrização que admite comportamentos crescentes ou decrescente ao longo do tempo, o modelo GTDL revela-se uma alternativa robusta para a análise de fenômenos com padrões de risco dinâmicos. Estudos como os de MacKenzie *et al.* (2003) e Blagojevic-Bucknall e MacKenzie (2004) compararam o modelo GTDL com modelos de fragilidade gama, analisando seu desempenho em dados multivariados e evidenciando sua aplicabilidade em diferentes contextos.

Uma limitação dos modelos existentes é a capacidade de acomodar a presença de indivíduos que nunca experimentarão o evento de interesse, mesmo sob longos períodos de observação. Esse fenômeno, conhecido como fração de cura, é frequentemente observado em estudos clínicos, nos quais uma parcela dos pacientes pode ser considerada curada, ou em engenharia, para componentes imunes a falhas. Para lidar com essa característica, uma classe de modelos conhecida como modelos defeituosos, cuja formulação remonta ao trabalho pioneiro de Berkson e Gage (1952) e foi formalmente desenvolvida por autores como Maller e Zhou (1996), tem se mostrado promissora, pois permite incorporar a fração de cura sem exigir pressuposições estritas sobre a distribuição dos tempos de sobrevivência.

Um exemplo clínico que ilustra a relevância de modelos com fração de cura é o estudo de pacientes com cirrose hepática, condição que representa o estágio terminal de diversas doenças hepáticas crônicas e constitui uma importante causa de mortalidade global. No Brasil, essas enfermidades são responsáveis por uma em cada 33 mortes, correspondendo a aproximadamente 3% de todos os óbitos registrados (de Oliveira e de Fátima Leite, 2024). A avaliação precisa do prognóstico desses pacientes é fundamental para subsidiar decisões clínicas e a alocação de recursos, sendo tradicionalmente realizada por meio de escores como o MELD (Model for End-Stage Liver Disease) e o Child-Pugh. No entanto, conforme demonstrado por Peng *et al.* (2016), esses sistemas apresentam limitações em capturar a heterogeneidade da doença, especialmente na identificação de subgrupos com sobrevida prolongada. Modelos de sobrevivência com fração de cura, ao incorporarem explicitamente a possibilidade de longos sobreviventes, oferecem uma alternativa promissora para melhorar a precisão prognóstica nesse contexto.

A necessidade de modelos mais precisos torna-se ainda mais premente diante do envelhecimento populacional e do aumento da prevalência de doenças hepáticas crônicas. No contexto do Sistema Único de Saúde (SUS), ferramentas prognósticas aprimoradas podem contribuir para uma alocação mais eficiente de recursos escassos, como vagas para transplante hepático. Além disso, a identificação de fatores associados à sobrevida prolongada pode revelar mecanismos protetores até então desconhecidos, abrindo novas linhas de investigação em hepatologia.

Neste trabalho, propõe-se uma nova distribuição de probabilidade com quatro parâmetros, denominada Família Estendida Logística Generalizada Dependente do Tempo (GTDEL). O objetivo é ampliar o modelo (GTDL), permitindo que se ajuste a uma gama mais diversificada de padrões de dados observados em contextos reais e calcular a fração de cura correspondente utilizando a ideia de modelo defeituoso. A extensão proposta busca superar limitações de modelos anteriores, oferecendo uma estrutura mais robusta para capturar a complexidade de fenômenos onde a variabilidade e a dinâmica temporal desempenham papéis cruciais.

São apresentadas algumas propriedades matemáticas do modelo, incluindo seus submodelos e o desenvolvimento do modelo de regressão correspondente. Adicionalmente, conduz-se um estudo de simulação via método de Monte Carlo, no qual diferentes cenários são avaliados a fim de verificar o desempenho dos estimadores obtidos por máxima verossimilhança.

Para demonstrar sua aplicabilidade prática, o modelo GTDEL é ajustado a quatro conjuntos de dados, sendo três compostos por dados reais de sobrevivência com censura e um

referente a dados de confiabilidade sem censura, evidenciando seu potencial para contribuir em análises aplicadas nas áreas da saúde e da engenharia.

1.1 Objetivos da dissertação

1.1.1 Objetivo Geral

Esta dissertação tem como principal objetivo propor a ampliação da família Logística Dependente do Tempo por meio da criação de uma nova distribuição de probabilidade com quatro parâmetros, denominada modelo Logístico Generalizado Estendido Dependente do Tempo (GTDEL), que é uma generalização do modelo de riscos não proporcional apresentado por Mackenzie (1996). O modelo GTDEL incorpora um novo parâmetro, permitindo maior flexibilidade para capturar a variabilidade dos dados e lidar com situações em que a suposição de proporcionalidade dos riscos não é válida.

1.1.2 Objetivos Específicos

Para alcançar o objetivo geral, estabeleceram-se os seguintes objetivos específicos:

- Avaliar a adequação do modelo GTDEL em diferentes cenários de dados, comparando-o com abordagens tradicionais.
- Desenvolver processos inferenciais para o modelo, incluindo estudos de simulação via Monte Carlo, com o objetivo de avaliar as propriedades assintóticas dos estimadores dos parâmetros.
- Explorar algumas propriedades matemáticas e estatísticas do modelo, destacando sua aplicabilidade em estudos de tempos de sobrevivência.
- Aplicar o modelo a quatro conjuntos de dados, sendo três compostos por dados reais de sobrevivência com censura e um por dados de confiabilidade sem censura, a fim de demonstrar sua eficiência prática em diferentes contextos.

O modelo proposto é fundamentado na análise de quatro conjuntos de dados reais, empregados em diferentes etapas do estudo. Os três primeiros foram utilizados para avaliar a adequação da distribuição proposta, enquanto o quarto foi aplicado no desenvolvimento e validação do modelo de regressão, permitindo uma análise mais abrangente de suas aplicações práticas.

Os conjuntos de dados utilizados neste trabalho foram selecionados com o intuito de avaliar a adequação e a aplicabilidade do modelo proposto em diferentes contextos. O primeiro conjunto refere-se a dados de resistência de fibras de vidro, coletados pelo National Physical Laboratory, na Inglaterra, e utilizados para descrever tempos de falha de materiais, permitindo analisar o desempenho do modelo em cenários de confiabilidade. O segundo conjunto compreende dados de malária, disponibilizados por Colosimo e Giolo (2021), referentes a um estudo experimental conduzido no Centro de Pesquisas René Rachou (Fiocruz, MG), com foco na avaliação da eficácia de imunização contra a doença.

O terceiro conjunto de dados corresponde às informações clínicas de pacientes atendidos na Creighton University School of Dentistry, entre agosto de 2007 e março de 2013, reunidas sob o nome *Teeth*. Esses dados, disponíveis no pacote `MST` da linguagem R, foram empregados na modelagem de sobrevivência dentária. Por fim, foram utilizados os dados de colangite biliar primária *pbc*, provenientes de um estudo realizado pela Mayo Clinic, nos Estados Unidos, entre 1974 e 1984. Esses dados estão disponíveis no pacote `survival` do R (R Core Team, 2024) e foram fundamentais para a aplicação e validação do modelo de regressão proposto.

1.2 Organização da dissertação

Esta dissertação está organizada em sete capítulos, estruturados de forma a proporcionar uma compreensão progressiva dos conceitos e do modelo proposto. O Capítulo 2 apresenta uma revisão abrangente sobre os principais conceitos da análise de Sobrevivência. O Capítulo 3 introduz o modelo GTDL, discutindo suas propriedades teóricas e limitações. No Capítulo 4, propõe-se a distribuição GTDEL, sendo apresentadas suas funções principais e uma análise detalhada de suas características.

O Capítulo 5 é dedicado à apresentação de quatro aplicações práticas do modelo proposto, com o objetivo de avaliar seu desempenho e sua flexibilidade frente a diferentes tipos de dados. A primeira aplicação, baseada em dados de resistência de fibras de vidro, envolve observações sem censura, permitindo avaliar o ajuste da distribuição GTDEL em um cenário mais controlado. A segunda e a terceira aplicações utilizam dados censurados de malária e de sobrevivência dentária (Teeth) que visam examinar a adequação da distribuição GTDEL em contextos biomédicos com censura à direita. A quarta aplicação, com os dados de colangite biliar primária (PBC), inclui a modelagem com covariáveis, permitindo validar o modelo de regressão GTDEL e explorar seu uso em situações reais permitindo avaliar o desempenho do modelo em diferentes contextos. O Capítulo 6 introduz o modelo de regressão GTDEL, explorando sua capacidade de incorporação de covariáveis. Por fim, o Capítulo 7 apresenta as conclusões do trabalho, sintetizando os principais resultados, apontando limitações e sugerindo direções para pesquisas futuras.

1.3 Suporte computacional

As análises estatísticas foram conduzidas no ambiente R (versão 4.4.1) (R Core Team, 2024), uma linguagem de programação especializada em computação estatística e geração de gráficos, originalmente desenvolvida por Ihaka e Gentleman (1996). Essa linguagem permite a implementação de técnicas estatísticas eficientes e precisas, consolidando-se como uma ferramenta de grande relevância no campo computacional numérico. A plataforma está disponível para download gratuito no endereço oficial: https://www.R-project.org/.

A elaboração deste documento foi realizada na plataforma Overleaf, que utiliza como base a distribuição TeX Live 2024 para compilação dos arquivos em LaTeX. Durante o desenvolvimento do trabalho, foram empregados diversos recursos computacionais, com destaque para a utilização de pacotes específicos da linguagem R.

Para a obtenção dos estimadores de máxima verossimilhança, utilizou-se o método de Broyden-Fletcher-Goldfarb-Shanno (BFGS), implementado pela função `optim(·, method = "BFGS")` da linguagem R (R Core Team, 2024). O modelo proposto foi ajustado a conjuntos de dados censurados, incluindo os dados de cirrose biliar primária (*pbc*), disponíveis no pacote `survival`, e os dados de sobrevivência dentária (*Teeth*), acessíveis no pacote `MST`.

Além disso, com o objetivo de comparação, aplicou-se o modelo de regressão de Cox por meio da função `coxph(·)`, também disponível no pacote `survival`. A suposição de riscos proporcionais foi avaliada com o teste baseado nos resíduos de Schoenfeld, implementado pela função `cox.zph()`.

Para a visualização gráfica dos resultados, foram utilizadas as bibliotecas `ggplot2` e `survminer`, que possibilitam a construção de gráficos de alta qualidade e personalizáveis, incluindo curvas de Kaplan-Meier e boxplots com facetas.

Capítulo 2

Análise de Sobrevivência

Neste capítulo, são apresentados os aspectos principais relacionados à Análise de Sobrevivência. Inicialmente, são abordados alguns conceitos fundamentais, seguidos pela introdução do estimador de Kaplan-Meier, amplamente utilizado para estimar funções de sobrevivência. Em seguida, são discutidos modelos probabilísticos comuns nessa área, bem como o método da máxima verossimilhança para a estimação dos parâmetros desses modelos. O modelo de riscos proporcionais de Cox, um dos mais empregados na literatura, é detalhado com suas propriedades e limitações. Por fim, é tratada a fração de cura em modelos defeituosos, uma abordagem relevante para estudos em que parte da população pode não experimentar o evento de interesse.

2.1 Conceitos básicos

Nas últimas décadas, a análise de sobrevivência tem se destacado como uma área em rápido crescimento dentro da estatística, impulsionada pelo avanço dos recursos computacionais e o aperfeiçoamento de técnicas estatísticas. Amplamente utilizada em campos como medicina, biologia, engenharia e demografia, essa abordagem foca no estudo do tempo até a ocorrência de um evento de interesse acontecer, como a morte de um paciente ou a falha de um componente. Esse tempo é conhecido como tempo de falha ou tempo de sobrevivência, representado pela variável aleatória T.

Uma variável aleatória é uma função do espaço amostral Ω no conjunto dos números reais, para a qual é possível calcular a probabilidade de ocorrência de seus valores. Em geral, são representadas por letras maiúsculas do alfabeto e sua classificação é feita de acordo com os valores que assumem (Magalhães, 2006). Nesta dissertação, são utilizadas variáveis aleatórias contínuas, que são definidas por apresentarem uma função não negativa f tal que para todo $x \in \mathbb{R}$:

$$F(t) = \int_{-\infty}^{x} f(u)du.$$

Uma característica essencial dessa análise é a censura, que ocorre quando o acompanhamento de uma observação é interrompido antes do evento de interesse acontecer, como no caso de um paciente que abandona o estudo ou de um componente removido antes de falhar.

A censura é um conceito fundamental na análise de sobrevivência e é representada pela variável aleatória C, que indica o tempo até a ocorrência de censura. Junto com o tempo de falha T, a variável C compõe a resposta observada no modelo. Em particular, o tempo observado para cada indivíduo é dado pelo mínimo entre o tempo de falha e o tempo de censura, ou seja, $t_i = min\{T_i, C_i\}, \forall i = 1, 2, ..., n$.

Para uma amostra aleatória de tamanho n, as observações são representadas como $(t_i, d_i, \boldsymbol{X}_i)$, em que t_i representa o tempo observado, que pode ser um tempo de falha ou de censura; d_i é o indicador de censura, que assume o valor 1 se t_i é um tempo de falha (ou seja, $T_i \leq C_i$) e 0 se t_i é um tempo censurado (ou seja, $T_i > C_i$) e $\boldsymbol{X}_i$ é o vetor de covariáveis, que pode incluir informações como sexo, idade, tratamento recebido, entre outras.

As censuras podem acontecer por várias razões, como, por exemplo, o término do experimento, o paciente ter se mudado para outra localidade, o paciente abandonar o tratamento, quebra do equipamento por motivo diferente do estudado, entre outros. Em estudos clínicos, diferentes tipos de censura podem ser observados, cada um com características específicas. A censura do tipo I ocorre quando o estudo é finalizado após um período previamente estabelecido, e, ao término desse intervalo, alguns indivíduos ainda não apresentaram o evento de interesse. Já a censura do tipo II está associada a estudos que se encerram assim que um número predefinido de indivíduos experimenta o evento em questão, independentemente do tempo decorrido. Por fim, a censura aleatória refere-se a situações em que um indivíduo é retirado do estudo antes da ocorrência do evento de interesse, sem que haja um critério fixo ou tempo preestabelecido, sendo, portanto, dependente de fatores diversos e não controlados.

Além dessa classificação associada ao planejamento do estudo, é possível caracterizar a censura segundo a forma como os dados censurados se manifestam. De acordo com Dey *et al.* (2020), os mecanismos podem ser descritos como censura à direita, quando o evento não ocorreu até o final do estudo; censura à esquerda, quando o evento ocorreu antes do início do estudo; e censura intervalar, que ocorre quando o evento de interesse não pode ser observado diretamente, mas se conhece o intervalo de tempo em que houve o desfecho.

O comportamento da variável aleatória contínua e não negativa $T \geq 0$, que representa o tempo de falha ou tempo de sobrevivência, pode ser representado por algumas funções matematicamente equivalentes, como por exemplo a função densidade de probabilidade (fdp), $f(t)$; a função de sobrevivência, $S(t)$ e a função de risco, $h(t)$.

A fdp é definida como sendo a probabilidade de um indivíduo falhar no intervalo de tempo $[t, t + \Delta t]$, por unidade de tempo e é dada por:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t},$$

em que a fdp possui as seguintes propriedades fundamentais: (i) $f(t) \geq 0$, $\forall t \in \mathbb{R}$, ou seja, a função densidade de probabilidade é sempre não negativa; e (ii) a integral da fdp sobre todo o domínio real é igual a 1, isto é,

$$\int_{-\infty}^{\infty} f(u)du = 1,$$

assegurando que a função represente uma distribuição de probabilidade válida.

A função de sobrevivência é uma das principais funções probabilísticas usadas para descrever estudos sobre tempo de sobrevivência e representa a probabilidade do evento de interesse não ocorrer em pelo menos t unidades de tempo. A função pode ser escrita como:

$$S(t) = P(T \geq t) = \int_{t}^{\infty} f(u)du = 1 - F(t),$$

em que a função de distribuição acumulada (fda), $F(t)$ é dada por:

$$F(t) = P(T \leq t) = 1 - P(T > t) = 1 - S(t).$$

O conhecimento da fda permite obter qualquer informação sobre a variável aleatória T. Mesmo que a variável só assuma valores num subconjunto dos reais, a função de distribuição é definida em toda reta e obedece as seguintes propriedades:

(i) $\lim_{t \to -\infty} F(t) = 0$ e $\lim_{t \to \infty} F(t) = 1$;

(ii) F é contínua à direita;

(iii) F é não decrescente.

As propriedades acima também poderiam ser usadas como definição da função de distribuição. James (2004), afirma que em termos mais abstratos, poderíamos dizer que toda função de $\mathbb{R}$ em $(0, 1)$, satisfazendo (i), (ii) e (iii), é fda de alguma variável aleatória.

Por outro lado, a função de risco ou taxa de falha, é utilizada para descrever o comportamento da variável tempo de sobrevivência e pode ser definida como o limite da probabilidade de um indivíduo falhar no intervalo de tempo $[t, t + \Delta t]$, dado que o indivíduo tenha sobrevivido até o instante t. Esta função, pode assumir formas constante, crescente, decrescente, unimodal ou em forma de banheira.

Para uma variável aleatória contínua T com $f(t)$ e $F(T)$, a função de risco é dada por:

$$h\left(t\right) = \lim_{\Delta t \to 0} \frac{P\left(T < t + \Delta t | T > t\right)}{\Delta t} = \frac{f(t)}{1 - F(t)}.$$

Na maioria dos casos, o tempo de sobrevivência está associado a causas do cotidiano que podem ser mais desafiadoras de representar matematicamente. Sendo assim, modelos paramétricos podem ser utilizados por sua comprovada adequação a situações práticas para modelar o tempo de sobrevivência até a ocorrência do evento de interesse. Entre os modelos, podemos citar o exponencial, o de Weibull, o gama e o gama generalizado.

A presença de dados censurados impõe um grau de dificuldade maior aos modelos paramétricos. Alternativamente, métodos não - paramétricos podem ser utilizados. Por exemplo, o estimador não- paramétrico de Kaplan Meier, para estimar a função de sobrevivência.

2.2 Estimador de Kaplan-Meier (KM)

Nesta Seção apresenta-se o estimador de Kaplan-Meier, que é uma técnica estatística não paramétrica amplamente utilizada para estimar a função de sobrevivência em estudos de tempo até o ocorrência de um evento de interesse, especialmente na presença de dados censurados. Devido à sua flexibilidade, essa ferramenta é aplicada em áreas como medicina, engenharia, ciências sociais e tem ganhando cada vez mais espaço em estudos de confiabilidade conforme destacado por Colosimo e Giolo (2021).

Proposto por Kaplan-Meier (1958), o método também conhecido como estimador limite-produto constitui uma adaptação da função de sobrevivência empírica, incorporando observações censuradas e permitindo, assim, estimativas mesmo com dados incompletos.

Na ausência de censuras, a função de sobrevivência empírica é definida por:

$$\widehat{S}(t) = \frac{\text{número de indivíduos que não falharam até o tempo } t}{\text{número total de indivíduos no estudo}},$$

em que $\widehat{S}(t)$ é uma função escada com degraus nos tempos observados de falha, de tamanho $1/n$, sendo n o tamanho da amostra.

Quando os dados incluem censura, a abordagem de Kaplan-Meier é ajustada para levar em consideração os indivíduos que saem do estudo antes da ocorrência do evento de interesse. O cálculo envolve o produto cumulativo das probabilidades de sobrevivência em cada tempo de falha observado, desconsiderando as censuras em cada intervalo.

Neste contexto, a metodologia baseia-se na estimativa de q_i, que representa a probabilidade de um indivíduo falhar no intervalo $[t_{i-1}, t_i)$, dado que ele sobreviveu até o tempo t_{i-1}.

A expressão de q_i, adaptada à definição de $\widehat{S}(t)$, é dada por:

$$\hat{q}_i = \frac{\text{número de falhas em } t_i}{\text{número de observações sob risco em } t_{i-1}},$$

para $i = 1, \ldots, k$.

A expressão geral do estimador de Kaplan-Meier para a função de sobrevivência é dada por:

$$\widehat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{e_i}{n_i}\right),$$

conforme descrito em Colosimo e Giolo (2021), em que t_i representa os tempos observados de falha, e_i corresponde ao número de eventos (falhas) no tempo t_i, e n_i indica o número de indivíduos em risco imediatamente antes do tempo t_i.

Dessa forma, a técnica oferece uma abordagem não paramétrica eficaz para estimar a função de sobrevivência, especialmente em contextos onde os dados censurados estão presentes. Contudo, em muitas situações práticas, é desejável complementar essa análise com modelos probabilísticos que permitam uma descrição mais detalhada dos mecanismos subjacentes aos dados. Esses modelos oferecem maior flexibilidade para incorporar hipóteses específicas sobre a distribuição dos tempos de falha e avaliar o impacto de variáveis explicativas na sobrevivência.

2.3 Modelos paramétricos em Análise de Sobrevivência

Os modelos probabilísticos são ferramentas fundamentais em análise de sobrevivência, permitindo a modelagem matemática do tempo até a ocorrência de um evento de interesse, como falha, morte ou recuperação.

Embora exista uma variedade de modelos probabilísticos aplicáveis à análise de sobrevivência, alguns se destacam por sua simplicidade e comprovada adequação a diferentes contextos práticos. Entre os mais utilizados, destacam-se o modelo exponencial, o de Weibull e gama.

Cada um desses modelos possui características específicas que os tornam adequados para diferentes tipos de dados e situações. Neste trabalho, exploraremos esses modelos, discutindo suas propriedades, aplicações e como eles podem ser utilizados para extrair insights valiosos a partir de dados de sobrevivência.

2.3.1 Distribuição Exponencial

A variável aleatória T segue a distribuição exponencial de parâmetro α ($\alpha > 0$) se tiver fdp dada por:

$$f(t;\alpha) = \alpha^{-1}\exp\{-t\alpha^{-1}\}, \quad t > 0$$

sendo representada por $T \sim \exp(\alpha)$.

O parâmetro α indica a taxa de ocorrência por unidade de medida, que pode ser tempo, distância, volume, entre outras. Esta distribuição destaca-se por sua simplicidade matemática e pela sua disposição uniparamétrica, sendo frequentemente utilizada em situações onde a taxa de falha é constante ao longo do tempo.

Um exemplo de aplicação da distribuição exponencial pode ser encontrado no trabalho de Gove (2017), que realiza um estudo demográfico aplicado a florestas com idades desiguais. Nesse estudo, os autores utilizam a distribuição exponencial para analisar a relação teórica entre o recrutamento, a mortalidade e o crescimento do diâmetro das árvores em diferentes setores. Essa aplicação ilustra a versatilidade da distribuição exponencial em contextos práticos, mesmo em áreas além da análise de sobrevivência tradicional.

Outro exemplo recente é o estudo de Ali *et al.* (2020), que aplica a distribuição exponencial na análise de confiabilidade de sistemas de engenharia. Os autores utilizam o modelo exponencial para modelar o tempo até falha de componentes eletrônicos, destacando sua utilidade em cenários onde a taxa de falha é constante e não depende da idade do componente. Esse trabalho reforça a relevância da distribuição exponencial em aplicações práticas modernas, especialmente em engenharia e tecnologia.

As funções de sobrevivência e de risco da distribuição exponencial são dadas por:

$$S(t;\alpha) = \exp\{-t\alpha^{-1}\} \quad e$$

$$h(t;\alpha) = \alpha^{-1}.$$

A Figura 1 mostra algumas formas das funções densidade de probabilidade, de sobrevivência e de risco da distribuição exponencial, sob diferentes valores do parâmetro α.

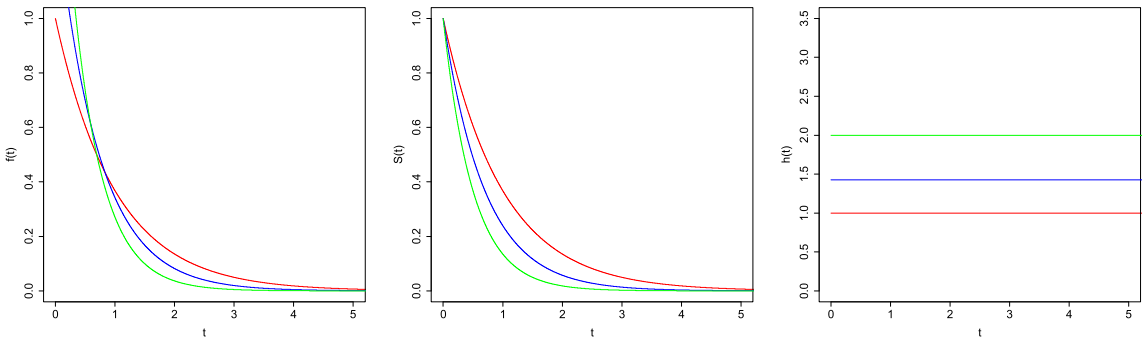


Figura 1: Função densidade de probabilidade (painel esquerdo), de sobrevivência (painel central) e de risco (painel direito) da distribuição exponencial para diferentes valores de $\alpha = 1,0$ (vermelho), $\alpha = 0,7$ (azul) e $\alpha = 0,5$ (verde).

2.3.2 Distribuição de Weibull

Proposta por Weibull (1951), a distribuição Weibull é bastante utilizada em casos de estudos biomédicos e industriais. Sua popularidade em aplicações práticas se deve ao fato de que sua fdp apresenta uma grande variedade de formas.

Uma variável aleatória T possui distribuição de Weibull com parâmetro α e β quando sua fdp é dada por:

$$f(t;\alpha,\beta) = \frac{\beta}{\alpha^{\beta}} t^{\beta-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}. \quad t > 0.$$

As funções de sobrevivência e de risco da distribuição weibull são definidas por:

$$S(t;\alpha,\beta) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\} \quad \text{e}$$

$$h(t;\alpha,\beta) = \frac{\beta}{\alpha^{\beta}} t^{\beta-1},$$

em que $\alpha > 0$ é o parâmetro de escala e $\beta > 0$ é o parâmetro de forma. Um caso particular da distribuição Weibull é a distribuição exponencial quando o parâmetro $\beta = 1$.

A Figura 2 revela as possíveis formas das funções densidade de probabilidade, de sobrevivência e de risco da distribuição Weibull, com $\alpha = 250$ e sob diferentes valores de β.

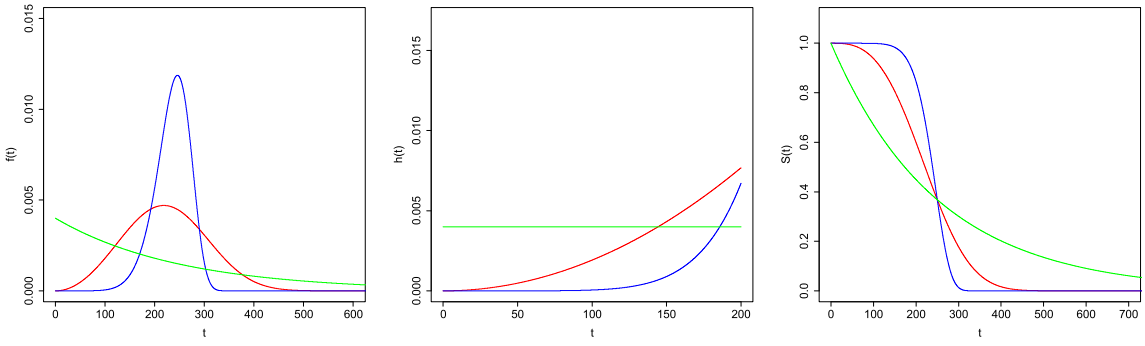


Figura 2: Função densidade de probabilidade (painel esquerdo), de sobrevivência (painel direito) e de risco (painel central) da distribuição de Weibull para diferentes vetores de parâmetros$\beta = 3,0$ (vermelho), $\beta = 1,0$ (verde) e $\beta = 8,0$ (azul).

2.3.3 Distribuição Gama

A distribuição gama foi formalizada como uma generalização de outras distribuições importantes, como a exponencial e a qui-quadrado (Casella e Berger, 2024). Seu nome deriva da função gama, um conceito fundamental na matemática, introduzido por Euler no século XVIII (Boyer e Merzbach, 2011).

A função Gama age como uma extensão do conceito de fatorial, ela é definida da seguinte forma:

$$\Gamma(\beta) = \int_0^\infty x^{\beta-1}\exp(-x)dx, \beta > 0,$$

para $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$.

Uma variável aleatória T possui distribuição Gama com parâmetro α e β quando sua fdp é dada por:

$$f(t;\alpha,\beta) = \frac{1}{\Gamma(\beta)\alpha^\beta}t^{\beta-1}\exp\left\{-\frac{t}{\alpha}\right\},$$

em que $\alpha > 0$ é o parâmetro de escala, $\beta > 0$ é o parâmetro de forma e representa-se $T \sim Gama\ (\alpha, \beta)$.

As funções de sobrevivência e de risco da distribuição gama são definidas por:

$$S(t;\alpha,\beta) = \int_t^\infty \frac{1}{\Gamma(\beta)\alpha^\beta}u^{\beta-1}\exp\left\{-\frac{u}{\alpha}\right\}du,$$

e

$$h(t;\alpha,\beta) = \frac{f(t;\alpha,\beta)}{S(t;\alpha,\beta)}.$$

A Figura 3 mostra algumas formas das funções densidade de probabilidade, de sobrevivência e de risco da distribuição gama, com $\alpha = 1$ e sob diferentes valores de β.

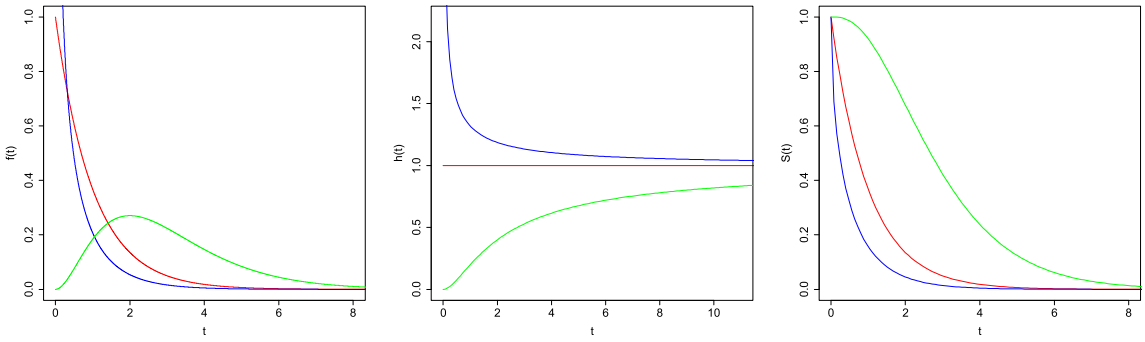


Figura 3: Função densidade de probabilidade (painel esquerdo), de risco (painel central) e de sobrevivência (painel direito) da distribuição gama para diferentes valores de $\beta = 1,0$ (vermelho), $\beta = 3$ (verde) e $\beta = 0,5$ (azul).

A distribuição Gama é amplamente utilizada na análise de dados assimétricos. Embora seu uso na modelagem de tempos de falha seja limitado, ela é bastante popular na análise de variáveis meteorológicas. Por exemplo, Longo *et al.* (2006) avaliaram as distribuições Gama e Log-Normal na estimativa de precipitações pluviais quinzenais no Estado do Paraná, utilizando dados diários de precipitação provenientes de 22 estações de medição. Os resultados indicaram

que a distribuição Gama apresentou melhor ajuste às condições pluviométricas da região. Já Shakil e Kibria (2009) empregaram a distribuição da combinação linear de variáveis Gama e Rayleigh para modelar dados de precipitação, demonstrando a qualidade do ajuste por meio do teste qui-quadrado de aderência.

Alguns modelos probabilísticos importantes são casos particulares da distribuição gama. A Tabela 2.1 apresenta os principais casos particulares da distribuição gama.

Tabela 2.1: Casos particulares da distribuição gama.

Distribuição	Parâmetros
Exponencial	$\alpha = \beta = 1,\ \beta > 0$
Qui-quadrado	$\alpha = \frac{n}{2},\ \beta = \frac{1}{2},\ n > 0$ inteiro
Erlang	$\alpha = n > 0,\ \beta > 0,\ n$ inteiro

Uma das características marcantes da distribuição gama é sua flexibilidade.

2.3.4 Distribuição gama generalizada

Introduzida por Stacy (1962), a distribuição gama generalizada estende a Gama padrão por meio de um terceiro parâmetro (τ), permitindo maior flexibilidade, e tem sido aplicada em diversos estudos. Por exemplo, Ramos (2014) utilizou a distribuição Gama Generalizada para analisar os dados de tempo de vida de 30 unidades provenientes de um experimento industrial, comparando os resultados obtidos com aqueles gerados pelas distribuições Weibull, Gama e Log-Normal. Já Castro *et al.* (2016), empregaram a distribuição para descrever a distribuição diamétrica de povoamentos de eucalipto, evidenciando sua versatilidade em diferentes contextos de modelagem.

Uma variável aleatória T possui distribuição gama generalizada com parâmetro $\alpha > 0$, $\beta > 0$ e $\tau > 0$, quando sua fdp é dada por:

$$f(t;\alpha,\beta,\tau) = \frac{\tau}{\alpha\Gamma(\beta)}\left(\frac{t}{\alpha}\right)^{\tau\beta-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\tau}\right\},$$

em que $\Gamma(\beta)$ é a função gama:

$$\Gamma(\beta) = \int_0^{\infty} w^{\beta-1}\exp\{-w\}dw.$$

Nesta parametrização, temos que α é o parâmetro de escala, enquanto β e τ são parâmetros de forma.

As funções de sobrevivência e de risco associadas à distribuição gama generalizada são dadas por:

$$S(t;\alpha,\beta,\tau) = 1 - \gamma_1\left[\beta, \left(\frac{t}{\alpha}\right)^{\tau}\right],$$

em que $\gamma_1(\beta, x)$ é a função gama incompleta normalizada, definida por:

$$\gamma_1(\beta, x) = \frac{\gamma(\beta, x)}{\Gamma(\beta)}.$$

Sendo $\gamma(\beta, x) = \int_0^x w^{\beta-1} \exp\{-w\} dw$ a função gama incompleta, que calcula a integral acumulada da função densidade gama até o ponto x. A função de risco é definida por:

$$h(t; \alpha, \beta, \tau) = \frac{f(t)}{S(t)} = \frac{t^{\tau\beta-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\tau}\right\}}{\int_0^{\infty} x^{\tau\beta-1} \exp\left\{-\left(\frac{x}{\alpha}\right)^{\tau}\right\} dx}.$$

Esta distribuição, apresenta flexibilidade também ao caracterizar vários modelos conhecidos ao restringir seus parâmetros.

A Tabela 2.2 apresenta os principais casos particulares da distribuição gama generalizada.

Tabela 2.2: Casos particulares da distribuição gama generalizada.

Distribuição	Parâmetros
Exponencial	$\beta = \tau = 1$
Weibull	$\beta = 1$
Gama	$\tau = 1$

A Figura 4 mostra algumas formas das funções densidade de probabilidade, de sobrevivência e de risco da distribuição gama generalizada, para o parâmetro $\alpha = 1$ e sob diferentes valores dos parâmetros β e τ.

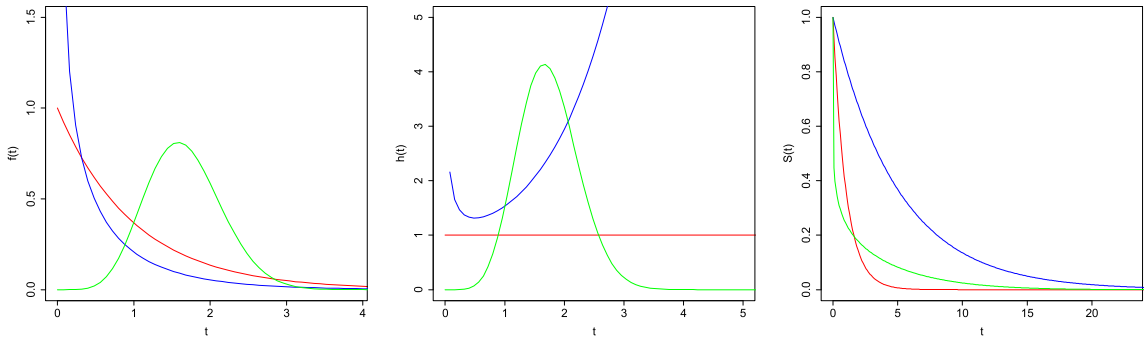


Figura 4: Função densidade de probabilidade (painel esquerdo), de sobrevivência (painel direito) e de risco (painel central) da distribuição gama generalizada para diferentes valores de $\boldsymbol{\theta}_1 = (\beta = 1{,}0;\ \tau = 1{,}0)^{\top}$(vermelho), $\boldsymbol{\theta}_2 = (\beta = 3{,}0;\ \tau = 2{,}0)^{\top}$(verde), $\boldsymbol{\theta}_3 = (\beta = 0{,}5;\ \tau = 1{,}0)^{\top}$(azul) .

Os parâmetros de um modelo probabilístico são as quantidades desconhecidas que se deseja estimar com base em uma amostra. Para isso, diversos métodos podem ser utilizados, sendo o método da máxima verossimilhança um dos mais aplicados na inferência estatística.

Esse método busca determinar os valores dos parâmetros que maximizam a função de verossimilhança, a qual expressa a probabilidade de a amostra observada ter sido gerada pelos valores escolhidos para os parâmetros do modelo.

2.4 Método da Máxima Verossimilhança

Existem diversos métodos de estimação conhecidos na literatura, sendo o método dos mínimos quadrados talvez o mais conhecido. No entanto, este método é inadequado para estudos de tempo de vida, principalmente devido à sua incapacidade de lidar com dados censurados durante o processo de estimação. Para esses casos, o método de máxima verossimilhança surge como uma alternativa apropriada. Ele não apenas incorpora as censuras nos cálculos, é simples de compreender e possui propriedades assintóticas ótimas, especialmente para grandes amostras. Para dados sem censuras, conforme em Bolfarine e Sandoval (2001) a função de verossimilhança para uma amostra aleatória de tamanho n é definida por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [f(t_i; \boldsymbol{\theta})].$$

Seja T_i, para $i = 1, 2, \ldots, n$, uma variável aleatória que denota o tempo de falha. Assumindo que $T_i's$ variáveis aleatórias independentes e identicamente distribuídas, temos que $L(\boldsymbol{\theta})$ na presença de dados censurados é dada por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [f(t_i; \boldsymbol{\theta})]^{d_i} [S(t_i; \boldsymbol{\theta})]^{1-d_i}, \tag{2.1}$$

em que d_i é o indicador de censura, com $d_i = 0$ se o tempo observado é censurado e $d_i = 1$ caso contrário. Como mostra a Equação (2.1), $f(t_i; \boldsymbol{\theta})$ representa a função densidade de probabilidade, e $S(t_i; \boldsymbol{\theta})$, a função de sobrevivência. A função de verossimilhança evidencia que cada observação contribui de forma distinta: para tempos censurados, a contribuição é dada pela função de sobrevivência; para tempos não censurados, a contribuição ocorre por meio da função densidade.

Outra forma de reescrever a função de verossimilhança é:

$$L(\boldsymbol{\theta}) \quad = \quad \prod_{i=1}^{n} [h(t_i; \boldsymbol{\theta})]^{d_i} \, S(t_i; \boldsymbol{\theta}),$$

onde $h(t_i; \boldsymbol{\theta})$ é a função de risco, uma razão entre $f(t_i; \boldsymbol{\theta})$ e $S(t_i; \boldsymbol{\theta})$.

Para facilitar a análise e o cálculo, é comum trabalhar com o logaritmo da função de verossimilhança, que transforma o produto em uma soma, sendo dado por:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} [d_i \log f(t_i; \boldsymbol{\theta}) + (1 - d_i) \log S(t_i; \boldsymbol{\theta})]. \tag{2.2}$$

Os Estimadores de Máxima Verossimilhança (EMV) são valores dos parâmetros $\boldsymbol{\theta}$ que maximizam a função de verossimilhança, com base nos dados observados. Esses estimadores são obtidos derivando $\ell(\boldsymbol{\theta})$ em relação ao vetor de parâmetros desconhecidos $\boldsymbol{\theta}$ e resolvendo o sistema de equações resultante. O método da máxima verossimilhança é amplamente utilizado em estatística devido às suas excelentes propriedades teóricas, especialmente quando o tamanho da amostra n é grande.

As propriedades assintóticas dos estimadores de máxima verossimilhança são essenciais para inferência estatística, permitindo construir intervalos de confiança, testar hipóteses e avaliar

a precisão dos estimadores.

Essas propriedades baseiam-se no fato de que, sob certas condições de regularidade, como a existência de derivadas e a não singularidade da matriz de informação de Fisher, ou seja, $det(I(\boldsymbol{\theta})) \neq 0$, o estimador $\widehat{\boldsymbol{\theta}}$ tem distribuição normal assintótica multivariada, expressa por:

$$\widehat{\boldsymbol{\theta}} \sim N_p(\boldsymbol{\theta}, I^{-1}(\boldsymbol{\theta})),$$

em que N_p denota a distribuição normal multivariada com p dimensões (onde p é o número de parâmetros em $\boldsymbol{\theta}$), $\boldsymbol{\theta}$ representa a média da distribuição normal multivariada, e $I^{-1}(\boldsymbol{\theta})$ é a inversa da matriz de informação de Fisher. Essa matriz corresponde à matriz de variâncias e covariâncias dos estimadores, sendo $I(\boldsymbol{\theta})$ definida como:

$$I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right],$$

onde, $\dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$ é a matriz Hessiana da log-verossimilhança em relação aos parâmetros $\boldsymbol{\theta}$, ou seja, a matriz composta pelas segundas derivadas parciais.

No entanto, o cálculo da matriz de informação de Fisher $I(\boldsymbol{\theta})$, não é matematicamente tratável devido à presença de censura nos dados. Nesse caso, pode-se utilizar, a matriz de informação observada, que é uma estimativa consistente de $I(\boldsymbol{\theta})$.

A matriz de informação observada é uma estimativa da matriz de informação de Fisher $I(\boldsymbol{\theta})$, que mede a quantidade de informação que os dados fornecem sobre os parâmetros $\boldsymbol{\theta}$. Enquanto a matriz de informação de Fisher é baseada no valor esperado das segundas derivadas da log-verossimilhança, a matriz de informação observada é calculada diretamente a partir das segundas derivadas da log-verossimilhança avaliadas no valor estimado dos parâmetros $\widehat{\boldsymbol{\theta}}$. Matematicamente, a matriz de informação observada é definida como:

$$-\ddot{\ell}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}},$$

em que $\ddot{\ell}(\boldsymbol{\theta}) =$ é a matriz Hessiana da função de log-verossimilhança $\ell(\boldsymbol{\theta})$, e $\widehat{\boldsymbol{\theta}}$ é o estimador de máxima verossimilhança dos parâmetros $\boldsymbol{\theta}$.

Adicionalmente, métodos numéricos, como Newton-Raphson e Escore de Fisher, são utilizados para encontrar estimadores de máxima verossimilhança. Esses métodos estão detalhados em trabalhos como o de Nocedal e Wright (2006), que são referências na área de otimização numérica. A abordagem proposta por ele é especialmente relevante para resolver problemas de otimização em funções complexas, como ocorre na estimativa de parâmetros em modelos de sobrevivência com censura.

2.4.1 Critérios de seleção de modelos

Os critérios de seleção de modelos, são particularmente úteis para identificar modelos que oferecem um compromisso ideal entre qualidade de ajuste e simplicidade. O princípio fundamental por trás desses critérios é a penalização de modelos excessivamente complexos, evitando o sobreajuste, que ocorre quando o modelo captura ruídos ou variações específicas da amostra ao invés de refletir padrões gerais dos dados. Essa penalização é baseada no número de parâmetros do modelo e também no tamanho da amostra, incentivando a escolha de modelos mais parcimoniosos em cenários com maior quantidade de dados. Esses critérios, amplamente adotados na literatura, fornecem uma abordagem sistemática para comparar modelos e são essenciais na prática estatística, principalmente em contextos de grande complexidade, como dados censura-

dos ou alta dimensionalidade.

A seguir, apresentamos dois dos critérios mais comuns, que foram utilizados neste trabalho:

1. Critério de Informação de Akaike (AIC), (Akaike, 1974):

   Este critério busca encontrar um equilíbrio entre o ajuste do modelo e sua complexidade, penalizando o aumento do número de parâmetros, que é definido como:

$$AIC = -2\ell(\boldsymbol{\theta}) + 2p.$$

2. Critério de Informação Bayesiano (BIC):

   Também chamado de critério de Schwarz, (Schwarz, 1978), sendo assim chamado porque Schwarz forneceu um argumento bayesiano para prová-lo. Similar ao AIC, o BIC é um critério de avaliação de modelos definido em termos de probabilidade a posteriori e também penaliza a complexidade do modelo, mas de forma mais severa, especialmente para amostras maiores. Este critério é definido como:

$$BIC = -2\ell(\boldsymbol{\theta}) + p\log(n),$$

em que $\ell(\boldsymbol{\theta})$ é o logaritmo da função de verossimilhança, p é o número de parâmetros do modelo e n o tamanho da amostra em estudo.

Esses critérios são ferramentas indispensáveis para a comparação de modelos, com a escolha geralmente recaindo sobre o modelo que apresenta os menores valores de AIC ou BIC. Isso significa que o modelo ideal é aquele que equilibra um bom ajuste aos dados com a simplicidade estrutural, evitando o sobreajuste. No presente trabalho, os critérios de seleção desempenharam um papel fundamental, permitindo a avaliação rigorosa de diferentes especificações de modelos com base em princípios estatísticos sólidos e orientando a escolha do modelo mais adequado às características dos dados analisados.

No contexto da análise de sobrevivência, a escolha de um modelo adequado torna-se ainda mais relevante, dado que é necessário lidar com a complexidade introduzida por dados censurados e explorar as relações entre covariáveis e tempos de sobrevivência.

Nesse cenário, o modelo de riscos proporcionais de Cox emerge como uma solução semiparamétrica amplamente utilizada. Ele combina simplicidade de interpretação e eficiência computacional, permitindo modelar dados censurados sem fazer suposições rígidas sobre a distribuição dos tempos de sobrevivência. A estrutura do modelo de Cox, baseada na hipótese de proporcionalidade dos riscos, facilita a análise e a identificação de padrões significativos nos dados, ao mesmo tempo em que aproveita as propriedades dos critérios de seleção para aprimorar a especificação do modelo.

2.5 Modelo de riscos proporcionais de Cox

Acompanhando a evolução da ciência estatística, a modelagem de dados clínicos passou por um avanço com a introdução do modelo de regressão de Cox, proposto por Cox (1972). Esse modelo também chamado de modelo de riscos proporcionais de Cox, é uma técnica estatística utilizada para a análise de sobrevivência, onde o objetivo principal é entender como diferentes variáveis independentes afetam o tempo até a ocorrência de um evento, como a morte de pacientes em um estudo clínico, falha de equipamentos em engenharia, ou qualquer evento cujo

tempo de ocorrência seja o foco de interesse. O modelo de Cox assume que a função de risco $h(t)$ é dada por:

$$h(t) = h_0(t)g(\boldsymbol{X}^\top \boldsymbol{\beta}),$$

em que $h_0(t)$ é a função de risco basal (não especificada); $\boldsymbol{X}$ representa o vetor de covariáveis; e $\boldsymbol{\beta}$ é o vetor de coeficientes a serem estimados.

Esse modelo semi-paramétrico, composto pelo produto de dois componentes, um não-paramétrico, $h_0(t)$ e um paramétrico $g(\boldsymbol{X}^\top \boldsymbol{\beta})$ que é uma função positiva, que assume valor 1 quando o seu argumento é zero. O componente paramétrico é frequentemente utilizado na seguinte forma multiplicativa:

$$g(\boldsymbol{X}^\top \boldsymbol{\beta}) = \exp\{\boldsymbol{X}^\top \boldsymbol{\beta}\} = \exp\left\{\beta_1 X_1 + \ldots + \beta_p X_p\right\},$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros associado às covariáveis. Sua popularidade decorre da interpretação direta dos coeficientes β.

Uma importante premissa do modelo de Cox é a proporcionalidade dos riscos, o que significa que a razão de risco entre dois indivíduos i e j deve permanecer constante ao longo do tempo, isto é, seja i e j dois indivíduos quaisquer, então:

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}}{h_0(t)\exp\{\boldsymbol{X}_j^\top \boldsymbol{\beta}\}} = \exp\left\{\boldsymbol{X}_i^\top \boldsymbol{\beta} - \boldsymbol{X}_j^\top \boldsymbol{\beta}\right\}.$$

A ausência explícita da variável tempo na razão implica que ela permanece constante independentemente do instante considerado, caracterizando a propriedade de riscos proporcionais. Este modelo assume que o vetor de covariáveis tem um efeito multiplicativo na função de risco. Isso implica que a estrutura do modelo impõe proporcionalidade entre as funções de risco de diferentes níveis de covariáveis, não permitindo que elas se cruzem ou dependam explicitamente do tempo. Quando essa suposição é violada, o modelo de Cox pode produzir estimativas enviesadas.

Dado que a suposição de proporcionalidade dos riscos é fundamental para a validade do modelo de Cox, torna-se essencial verificar se essa condição é atendida antes de interpretar os resultados da regressão. Caso contrário, as estimativas obtidas podem ser enviesadas, comprometendo a inferência estatística e a validade das conclusões do estudo.

Diversos métodos têm sido desenvolvidos para avaliar a proporcionalidade dos riscos, incluindo abordagens gráficas e testes estatísticos baseados em resíduos. Por exemplo, métodos visuais, como o gráfico de Nelson-Aalen, permitem uma inspeção exploratória do comportamento das funções de risco, enquanto abordagens analíticas, como o teste baseado nos resíduos de Schoenfeld, fornecem uma avaliação estatística mais formal dessa suposição.

2.5.1 Verificação de proporcionalidade

A suposição de riscos proporcionais é um dos pilares fundamentais do modelo de Cox. No entanto, em alguns cenários, essa suposição pode ser violada, comprometendo a validade das inferências realizadas com base no modelo.

Para lidar com essa questão, diversos métodos foram propostos na literatura para verificar a proporcionalidade dos riscos. Entre eles, destacam-se os métodos gráficos que fornecem uma visualização intuitiva das possíveis violações e os métodos analíticos, como o teste baseado nos resíduos de Schoenfeld (1982). Neste trabalho, serão utilizados ambos os métodos para uma avaliação abrangente das possíveis violações.

O método gráfico consiste em avaliar visualmente a proporcionalidade dos riscos com base na estimativa da função de risco acumulado, $H(t)$, ajustada para diferentes grupos de uma

covariável.

Para variáveis contínuas, essa divisão é feita considerando a mediana como ponto de separação, classificando os indivíduos em dois grupos: aqueles com valores abaixo e aqueles com valores acima da mediana.

A construção do gráfico segue os seguintes passos:

1. Divisão dos dados:

   Os indivíduos são agrupados em dois conjuntos com base na mediana da covariável de interesse, distinguindo aqueles com valores menores ou iguais à mediana daqueles com valores superiores.

2. Estimativa de $H(t)$:

   A função de risco acumulado $H(t)$ é estimada separadamente para cada grupo, utilizando o estimador de Nelson-Aalen-Breslow (Colosimo e Giolo, 2021). Esse estimador é adequado porque incorpora dados censurados e fornece uma estimativa não paramétrica da função de risco acumulado.

3. Construção do gráfico:

   O gráfico é gerado ao se plotar o logaritmo da estimativa de $H(t)$ em função do tempo t ou do logaritmo do tempo $\log(t)$. As curvas são traçadas separadamente para os indivíduos com valores acima e abaixo da mediana da covariável analisada.

4. Interpretação do gráfico:

   A proporcionalidade dos riscos é avaliada pela comparação das curvas. Se forem aproximadamente paralelas, não há indícios de violação da suposição de riscos proporcionais. No entanto, se houver divergência entre as curvas ao longo do tempo, isso sugere que os riscos não são proporcionais para a covariável em questão, indicando a necessidade de considerar abordagens alternativas no modelo de análise de sobrevivência.

Por outro lado, o teste de proporcionalidade dos riscos de Schoenfeld, implementado pela função `cox.zph(.)` no software R, foi o método analítico utilizado para verificar a adequação do modelo de Cox aos dados. Esse teste avalia se os coeficientes de regressão permanecem constantes ao longo do tempo, uma condição fundamental para a validade do modelo de riscos proporcionais.

A análise é baseada na correlação entre os resíduos de Schoenfeld e o tempo, sob a hipótese nula de que não há relação entre eles, ou seja, de que os coeficientes não variam temporalmente.

Para cada variável incluída no modelo, é calculada uma estatística qui-quadrado que avalia a relação entre os resíduos e o tempo. Se o p-valor associado for menor que 0,05 ($p < 0,05$), significa que os efeitos das covariáveis podem variar ao longo do tempo. Para lidar com essa violação, abordagens alternativas, como modelos com termos dependentes do tempo ou estratificados, podem ser utilizadas.

Além da análise individual de cada variável, o teste também pode ser aplicado de forma global, avaliando se o modelo completo - considerando todas as covariáveis simultaneamente - apresenta indícios de violação da proporcionalidade dos riscos. Caso o teste indique violação para uma ou mais covariáveis, pode ser necessário adotar estratégias alternativas, como a inclusão de termos dependentes do tempo ou o uso de modelos que relaxam a suposição de riscos proporcionais.

Diferentemente dos métodos gráficos, que dependem da interpretação visual das curvas, o teste analítico fornece uma medida estatística formal e objetiva, reduzindo a subjetividade na avaliação da proporcionalidade dos riscos. Dessa forma, sua aplicação é fundamental para garantir a validade das inferências feitas com base no modelo de Cox.

Neste trabalho, os resultados desse teste foram empregados para assegurar que os modelos ajustados fossem adequados aos bancos de dados analisados, permitindo estimativas mais robustas e interpretações confiáveis sobre os fatores associados ao tempo de sobrevivência.

2.5.2 Método da Máxima Verossimilhança Parcial

O método da máxima verossimilhança parcial, proposto por Cox, é uma abordagem estatística usada para estimar os parâmetros do modelo de riscos proporcionais. A verossimilhança parcial é construída considerando apenas a ordem dos tempos de eventos, e não os tempos exatos.

Suponha que nos tempos $t_1, \ldots, t_k$ existam k falhas diferentes numa amostra com n indivíduos. A probabilidade de falha da i- ésima observação no tempo t_i é dada pela razão entre o risco do indivíduo i e a soma dos riscos de todos os indivíduos em risco no tempo t_i. A Função de Verossimilhança Parcial é o produto dessas probabilidades condicionais para todos os tempos de evento observados. A função de verossimilhança parcial é expressa por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}}{\sum\limits_{j \in R(t_i)} \exp\{\boldsymbol{X}_j^\top \boldsymbol{\beta}\}} = \prod_{i=1}^{n} \left(\frac{\exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}}{\sum\limits_{j \in R(t_i)} \exp\{\boldsymbol{X}_j^\top \boldsymbol{\beta}\}} \right)^{e_i}, \tag{2.3}$$

em que k representa o número de tempos de eventos distintos observados, t_i denota o tempo do i-ésimo evento, e $\boldsymbol{X}_i$ corresponde ao vetor de covariáveis do indivíduo que sofreu o evento nesse tempo. O conjunto $R(t_i)$ indica o risk set, ou seja, o conjunto de indivíduos que ainda não sofreram o evento e que também não foram censurados até esse instante. Por fim, e_i é uma variável indicadora de falha que assume valor 1 se o evento ocorreu no tempo t_i e 0 caso contrário.

A Função (2.3) é chamada de parcial porque elimina a dependência da função de risco basal $h_0(t)$, focando apenas na estimação dos coeficientes $\boldsymbol{\beta}$. As estimativas de $\boldsymbol{\beta}$ que maximizam $L(\boldsymbol{\beta})$ são obtidas por meio da resolução do sistema de equações definido por $U(\boldsymbol{\beta}) = 0$, em que $U(\boldsymbol{\beta})$ é o vetor escore resultante do cálculo da primeira derivada do logaritmo da função de verossimilhança, $\ell(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta}))$.

O método consiste em maximizar a Função (2.3) para encontrar os valores de $\boldsymbol{\beta}$ que melhor explicam os dados observados. Isso é feito resolvendo o sistema de equações definido pelo vetor escore $U(\boldsymbol{\beta}) = \mathbf{0}$, onde:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i \left[\boldsymbol{X}_i - \frac{\sum_{j \in R(t_i)} \boldsymbol{X}_j \exp(\boldsymbol{X}_j^\top \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{X}_j^\top \boldsymbol{\beta})} \right],$$

sendo $\boldsymbol{X}_i$ o vetor de covariáveis do indivíduo i.

A solução desse sistema fornece as estimativas de $\boldsymbol{\beta}$. A Função (2.3) assume que os tempos de sobrevivência são contínuos e, portanto, não considera a possibilidade de empates nos tempos observados. No entanto, em aplicações práticas, é comum que ocorram empates, isto é, situações em que dois ou mais indivíduos apresentam o mesmo tempo de evento (falha ou censura). A presença de empates complica o cálculo da verossimilhança parcial, pois a ordem exata em que os eventos ocorreram não é conhecida.

Quando há empates em k tempos distintos, a formulação original da Função de Verossimilhança Parcial precisa ser ajustada para incorporar essas observações simultâneas.

A Função (2.3), ajustada para lidar com empates nos tempos de falha, é expressa por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{S}_i^\top \boldsymbol{\beta})}{\left[\sum_{j \in R(t_i)} \exp(\boldsymbol{X}_j^\top \boldsymbol{\beta})\right]^{m_i}},$$

em que $\boldsymbol{S}_i$ representa a soma dos vetores de covariáveis associados às observações empatadas no tempo t_i, m_i é o número de empates nesse instante, e $R(t_i)$ corresponde ao conjunto de indivíduos em risco no tempo t_i, isto é, aqueles que ainda não sofreram o evento de interesse nem foram censurados antes desse tempo. Nesse caso, m_i é usado para ajustar a contribuição dos empates no cálculo da verossimilhança. Como múltiplos indivíduos sofrem o evento no mesmo tempo, a contribuição conjunta desses indivíduos é incorporada na função por meio de m_i .

Essa formulação ajustada leva em conta a contribuição conjunta de todas as observações empatadas em cada tempo t_i, permitindo que a verossimilhança parcial seja calculada de forma apropriada mesmo na presença de empates.

Em resumo, embora a função de verossimilhança parcial seja uma ferramenta poderosa, ela tem limitações em cenários com empates ou quando uma fração da população é imune ao evento de interesse. Nesses casos, a aproximação original pode não ser adequada. Além disso, o modelo de Cox não é apropriado quando há uma fração de cura, pois ele pressupõe que todos os indivíduos estão em risco até o final do estudo.

Para lidar com a presença de uma fração de cura, modelos defeituosos podem ser utilizados. Esses modelos permitem a inclusão de uma fração de indivíduos que nunca experimentarão o evento, tornando a modelagem mais realista em diversas áreas de aplicação. A formulação desses modelos baseia-se no conceito de distribuição de probabilidade defeituosa, que é fundamental para lidar com a fração de cura.

2.6 Modelo Defeituoso

Nesta Seção, é apresentado o conceito de distribuição de probabilidade defeituosa, que é utilizada para o cálculo da fração de curados da população. O método mais comumente utilizado na literatura para modelagem de fração de cura é o modelo de mistura padrão, inicialmente proposto por Boag (1949) e Berkson e Gage (1952), descrito por:

$$S(t) = p + (1-p)S_0(t), \tag{2.4}$$

onde $S_0(t)$ é uma função de sobrevivência própria, isto é,

$$\lim_{t \to +\infty} S_0(t) = 0.$$

Esta abordagem foi posteriormente expandida por Farewell (1982), demonstrando a importância de considerar subpopulações com diferentes perfis de risco. Esses modelos têm sido aplicados com sucesso em oncologia (Rodrigues *et al.*, 2009) e, em doenças crônicas.

Diferentemente desses modelos de mistura, as distribuições defeituosas oferecem uma alternativa paramétrica direta para estimar a fração de cura, sem exigir a especificação explícita de $S_0(t)$. Essa abordagem revolucionou a análise de sobrevivência ao capturar heterogeneidades populacionais por meio de propriedades intrínsecas da distribuição, como discutido em Balka *et al.* (2009).

Uma distribuição é chamada de defeituosa se a integral de sua função de densidade não resultar em 1, mas em um valor $\rho \in (0, 1)$, quando o domínio dos parâmetros é alterado (Rocha *et al.*, 2016), isto é:

$$\int_0^\infty f(t)dt = \rho < 1.$$

Nesse contexto, ρ é interpretado como a fração curada, ou seja, a proporção de indivíduos que nunca experimentarão o evento de interesse. Por outro lado, a fração faltante, $1 - \rho$, indica a probabilidade de que o evento ocorra para os indivíduos suscetíveis ao longo do tempo.

A definição de distribuição de probabilidade defeituosa, viola o segundo requisito fundamental das funções densidade de probabilidade, que exige que a integral da função sobre todo o domínio seja igual a 1. Essa violação compromete o uso do termo *função de distribuição de probabilidade*, pois contradiz os fundamentos matemáticos que caracterizam tal conceito. Além disso, a denominação *defeituosa* pode gerar confusão, uma vez que sugere que a função ainda pertence ao domínio das probabilidades.

A distribuição de probabilidade defeituosa é um conceito abordado no contexto de análise de sobrevivência, onde parte da população sob estudo nunca experimenta o evento de interesse, como uma falha, morte ou recaída de uma doença. Esse grupo de indivíduos que "não falha" é conhecido como fração de cura. Nesse contexto, o modelo assume que uma parcela da população é "curada"ou "imune"ao evento de interesse, de forma que o evento nunca ocorrerá para esses indivíduos, independentemente do período de observação.

Essas distribuições permitem captar a dualidade entre os indivíduos suscetíveis e curados, sendo particularmente úteis para modelar a sobrevivência em longo prazo e estimar a fração curada de maneira implícita, sem a necessidade de um modelo de mistura explícito.

Um exemplo clássico na literatura é o modelo GTDL, proposto por Mackenzie (1996), que incorpora a fração de cura que apresenta diferentes formatos para as funções de risco e de sobrevivência. Quando usado como um modelo para fração de cura, a proporção da população que é imune ρ, é obtida calculando o limite da função de sobrevivência com os parâmetros estimados.

$$\lim_{t\to+\infty} S\left(t\right) = \rho > 0,$$

em que $\rho \in (0, 1)$.

Entre as distribuições conhecidas na literatura que permitem a modelagem de fração de cura, estão a distribuição de Gompertz proposta por Gompertz (1825), e a distribuição Gaussiana inversa proposta por Whttmore (1979).

Diversos estudos destacam a aplicação dessas distribuições em diferentes contextos. Por exemplo, Rocha *et al.* (2016) utilizaram a classe de distribuições de Marshall–Olkin para generalizar a distribuição de Gompertz, criando novas distribuições defeituosas. Essas distribuições foram aplicadas a três conjuntos de dados reais, demonstrando a eficácia do modelo em cenários variados e evidenciando sua flexibilidade. Além disso, Santos *et al.* (2017) propuseram uma abordagem bayesiana para o modelo de Gompertz defeituoso, comparando-o com o modelo baseado no método de máxima verossimilhança. Por sua vez, Scudilio *et al.* (2019), desenvolveram um modelo defeituoso baseado na distribuição Gama, induzido por um termo de fragilidade, para modelar a proporção de curados em uma população. Neste estudo, as distribuições Gompertz e Gaussiana inversa defeituosas foram empregadas como funções base para ilustrar a definição de um modelo defeituoso.

Uma das principais vantagens dos modelos defeituosos, é que a fração de cura não precisa ser assumida previamente; ela emerge naturalmente quando os parâmetros assumem determinados valores fora da faixa esperada. Além disso, a presença e a significância dessa fração podem

ser avaliadas estatisticamente por meio dos parâmetros associados à estrutura defeituosa do modelo.

2.6.1 A Distribuição Gompertz defeituosa

Originalmente desenvolvida para modelar a mortalidade, a distribuição de Gompertz proposta por Gompertz (1825), encontra aplicações em diversas áreas, como atuária, biologia e estudos demográficos. Além disso, é amplamente utilizada para modelar dados de sobrevivência em diferentes contextos do conhecimento (Gieser *et al.*, 1998).

Trata-se de uma distribuição de probabilidade que descreve o risco de eventos que aumentam exponencialmente com o tempo, sendo particularmente útil para representar padrões de risco crescentes em populações ou sistemas dinâmicos.

A fdp da distribuição de Gompertz é expressa como:

$$f(t) = b\exp(at)\exp\left\{-\frac{b}{a}(\exp(at)-1)\right\},$$

para $a \in \mathbb{R}$, $b > 0$ e $t > 0$.

Nessa parametrização, a é o parâmetro de forma, enquanto b é o parâmetro de locação. A distribuição também pode admitir valores negativos para o parâmetro de forma, $(a < 0)$. Nesse caso, quando o parâmetro a assume valores negativos, a distribuição de Gompertz se torna imprópria. Os parâmetros que, ao alterar seus domínios, geram distribuições defeituosas são chamados de parâmetros defeituosos.

A função de sobrevivência, definida como,

$$S(t) = \exp\left\{-\frac{b}{a}(\exp(at)-1)\right\},$$

representa a probabilidade de um indivíduo sobreviver até o tempo t ou além dele. No contexto em que $a < 0$, a função de sobrevivência não converge para zero quando $t \to \infty$, mas sim para um valor positivo. Esse valor corresponde à proporção de indivíduos que nunca experimentam o evento de interesse, ou seja, a fração de curados na população.

A proporção de curados ρ é obtida ao calcular o limite da função de sobrevivência $S(t)$ quando $t \to \infty$, dado por:

$$\begin{aligned}
\rho &= \lim_{t\to+\infty} S(t) \\
&= \lim_{t\to+\infty} \exp\left\{-\frac{b}{a}(\exp(at)-1)\right\} \\
&= \exp\left\{-\frac{b}{a}\right\} \in (0,1).
\end{aligned}$$

A função de risco da Gompertz tem a forma:

$$h(t) = a \exp\{bt\}.$$

Aqui, a e b são parâmetros que controlam o nível e o crescimento exponencial do risco ao longo do tempo, onde o risco aumenta exponencialmente à medida que o tempo cresce.

A Figura 5 apresenta o comportamento das funções densidade de probabilidade, de sobrevivência e risco para a distribuição de Gompertz no caso em que $a < 0$, considerando os valores dos parâmetros $(-1, 1)$(vermelho), $(-2, 1)$(verde) e $(-1, 2)$(azul).

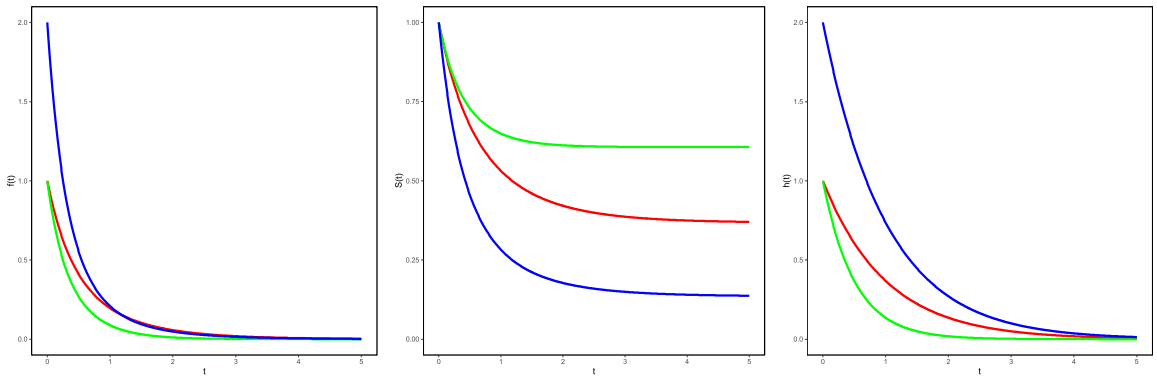


Figura 5: Função densidade de probabilidade (painel esquerdo), de sobrevivência (painel central) e de risco (painel direito) da distribuição Gompertz defeituosa para diferentes valores dos parâmetros.

2.6.2 A Distribuição Gaussiana Inversa defeituosa

A distribuição Gaussiana inversa, está intimamente ligada aos processos estocásticos e, surge como o primeiro tempo de passagem de um processo de Wiener. Um processo de Wiener é um exemplo de um processo estocástico contínuo, que modela o comportamento de uma partícula que se move de maneira aleatória no tempo. Esse problema é central na teoria dos processos estocásticos, tendo grande relevância em diversas áreas, como finanças, física e biologia. Além disso, Balka *et al.* (2009), Lee e Whitmore (2006) observaram seu potencial como modelos para a taxa de cura.

A distribuição Gaussiana Inversa possui fdp dada por:

$$f(t) = \frac{1}{\sqrt{2b\pi t^3}} \exp\left\{-\frac{1}{2bt}(1-at)^2\right\},$$

para $a \in \mathbb{R}$, $b > 0$ e $t > 0$. Para $a < 0$, temos um modelo defeituoso.

Uma função de sobrevivência definida como,

$$S(t) = 1 - \left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right],$$

onde $\Phi(\cdot)$ denota a função de distribuição cumulativa de uma variável aleatória normal padrão.

A fração de cura é calculada como o limite da função de sobrevivência,

$$\begin{aligned}\rho &= \lim_{t\to+\infty} S(t)\\ &= 1-\left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right]\\ &= 1-e^{2a/b}\in(0,1).\end{aligned}$$

A função de risco da distribuição Gaussiana Inversa é dada por:

$$h(t)=\frac{\frac{1}{\sqrt{2b\pi t^3}}\exp\left\{-\frac{1}{2bt}(1-at)^2\right\}}{1-\left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right)+e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right]}.$$

A Figura 6 mostra algumas formas das funções densidade de probabilidade, de sobrevivência e risco da distribuição Gaussiana Inversa Defeituosa, considerando os valores dos parâmetros $(-1;1)$(vermelho), $(-2;2)$(verde) e $(-1;2)$ (azul) respectivamente.

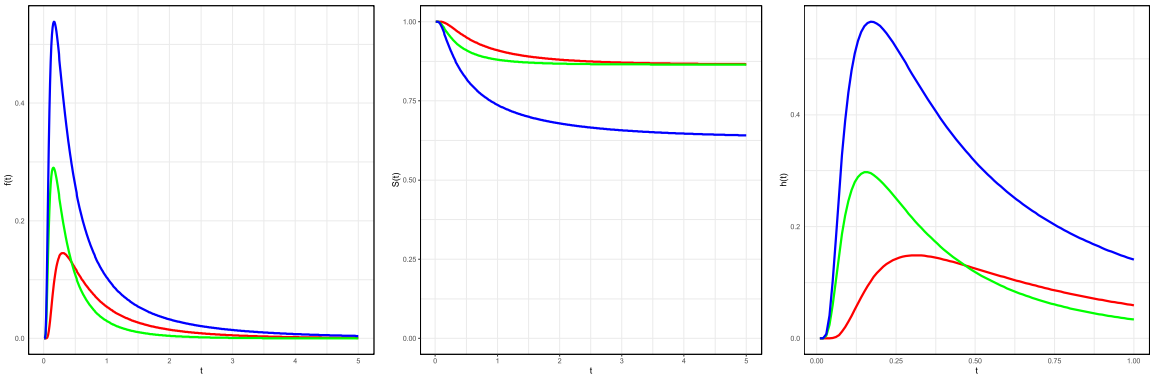


Figura 6: Função densidade de probabilidade (painel esquerdo), de sobrevivência (painel central) e de risco (painel direito) da distribuição Gaussiana Inversa defeituosa para diferentes valores dos parâmetros.

Este capítulo apresentou os fundamentos da Análise de Sobrevivência e destacou a necessidade de modelos mais flexíveis, para lidar com cenários em que há fração de cura. Além disso, discutimos a importância da verificação da proporcionalidade dos riscos e os critérios estatísticos para seleção do modelo mais adequado.

Capítulo 3

Modelo Logístico Generalizado Dependente do Tempo (GTDL)

Neste capítulo, apresenta-se o modelo GTDL, exploram-se suas principais funções, propriedades e suas características essenciais. Em particular, detalha-se o modelo GTDL defeituoso, enfatizando sua aplicação em cenários de longa duração. Por fim, apresenta-se o modelo de regressão GTDL, destacando sua formulação e potencial para análise de dados de sobrevivência.

3.1 O Modelo Probabilístico GTDL

Ao estudar tempos de sobrevivência de pacientes em um determinado experimento clínico, a modelagem via função de risco destaca-se como abordagem fundamental. O modelo de regressão de riscos proporcionais de Cox (1972), pioneiro na incorporação de covariáveis na análise, é amplamente reconhecido por sua simplicidade e versatilidade, conforme destacado por (Colosimo e Giolo, 2021). Entretanto, como discutido no Capítulo 2, o modelo de Cox assume riscos proporcionais, o que pode representar uma limitação em muitos contextos práticos, já que, em determinadas situações, a razão dos riscos pode variar ao longo do tempo, comprometendo a adequação do modelo.

O modelo GTDL (Generalized Time-Dependent Logistic), proposto por Mackenzie (1996), supera essa limitação ao incorporar uma estrutura dependente do tempo por meio de uma função logística modificada, tornando-se um modelo de risco não proporcionais que serve como uma alternativa viável para análises de dados de sobrevivência com padrões de risco dinâmicos, oferecendo uma alternativa robusta ao modelo de Cox.

Diversos estudos expandiram o modelo GTDL, por exemplo, Al-Tawarah e MacKenzie (2003), aplicaram o modelo GTDL como um modelo de sobrevivência não proporcional aos riscos (non-PH) com função de risco logístico, para analisar dados de ensaios clínicos longitudinais em que o tempo exato do evento de interesse era desconhecido devido à censura intervalar; Blagojevic-Bucknall e MacKenzie (2004), compararam a performance do modelo GTDL com e sem o termo de fragilidade para dados de sobrevivência multivariados; Milani (2011), estudou do modelo de risco logístico generalizado dependente do tempo com fragilidade; Louzada-Neto *et al.* (2010), a partir do enfoque Bayesiano apresentaram procedimentos inferenciais para o modelo GTDL; Louzada-Neto *et al.* (2011) apresentaram um estudo para estimar o intervalo de confiança para os parâmetros quando as amostras são pequenas; Louzada *et al.* (2020) aplicaram o modelo GTDL em válvulas de segurança do fundo de poços de petróleo *offshore*. Mais recentemente, Oliveira *et al.* (2023) empregaram resíduos como ferramenta de diagnóstico para avaliar a adequabilidade do modelo GTDL a dados reais de pacientes com câncer de pulmão em estágio avançado.

A análise de sobrevivência com dados que violam a suposição de riscos proporcionais também tem sido abordada por diversos autores. Clayton (1978) propôs modelos multivariados

para tabelas de vida, Kalbfleisch e Prentice (2011) desenvolveram o modelo de falha acelerada, e Hougaard (1984) tratou de distribuições que descrevem a heterogeneidade.

Seja T uma variável aleatória contínua e não negativa, que representa o tempo até a ocorrência do evento de interesse. Suponha que T possui distribuição GTDL, parametrizada por $\boldsymbol{\psi} = (\lambda, \alpha, \gamma)^\top$, com $\lambda > 0$, $\alpha > 0$, $\gamma \in \mathbb{R}$. Essa distribuição é caracterizada pelas funções de distribuição acumulada, densidade de probabilidade, e função de sobrevivência. A função de risco acumulada, por sua vez, é apresentada conforme MacKenzie e Peng (2014), e a função de risco é definida a partir da razão entre a densidade e a função de sobrevivência, conforme descrito a seguir:

$$F(t;\boldsymbol{\psi}) = 1 - \left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}, \tag{3.1}$$

$$f(t;\boldsymbol{\psi}) = \lambda\left(\frac{\exp(\alpha t+\gamma)}{1+\exp(\alpha t+\gamma)}\right)\left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}, \tag{3.2}$$

$$S(t;\boldsymbol{\psi}) = \left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}},$$

$$H(t;\boldsymbol{\psi}) = \int_0^t h(s)ds = \frac{\lambda}{\alpha}\log\left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right) \; e$$

$$h(t;\boldsymbol{\psi}) = \lambda\left(\frac{\exp(\alpha t+\gamma)}{1+\exp(\alpha t+\gamma)}\right). \tag{3.3}$$

A função de risco do modelo GTDL apresenta um comportamento monotônico que é diretamente influenciado pelo valor do parâmetro α. Esse parâmetro controla a variação do risco ao longo do tempo, resultando em diferentes dinâmicas de falha. Conforme discutido por Mackenzie (1996), a função de risco exibe uma forte dependência temporal, resultando em curvas com formato sigmoidal. Esse comportamento destaca a versatilidade do modelo para capturar padrões de falha que não seguem a suposição tradicional de riscos proporcionais.

O comportamento monotônico pode ser caracterizado da seguinte maneira: (i) quando $\alpha > 0$, a função de risco é crescente, indicando que a taxa de falha aumenta ao longo do tempo. Esse comportamento é típico de processos onde o risco de falha se intensifica à medida que o tempo avança; (ii) no caso em que $\alpha = 0$, a função de risco permanece constante ao longo do tempo, caracterizando um modelo de riscos proporcionais com função de risco basal exponencial. Nesse caso, a taxa de falha não é afetada pelo tempo, refletindo um cenário onde a probabilidade de falha é independente da duração da exposição ao risco.

A Figura 1 mostra algumas formas da função de risco do modelo GTDL para diferentes valores do paramétricos, evidenciando tanto a dependência temporal quanto o comportamento monotônico do modelo.

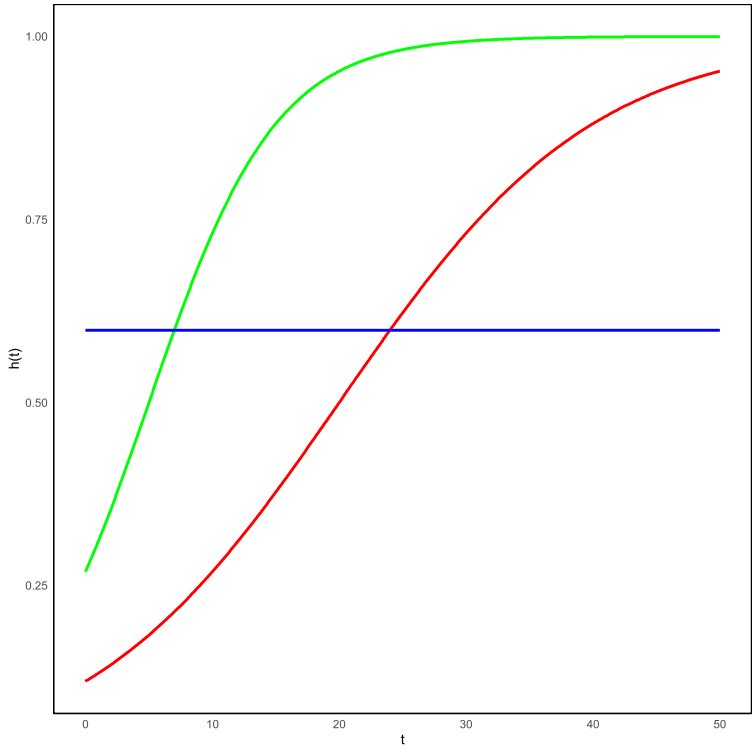


Figura 1: Formas da função de risco da distribuição GTDL para diferentes vetores de parâmetros: $\boldsymbol{\psi}_1 = (\lambda = 1{,}0;\ \alpha = 0{,}1;\ \gamma = 2{,}0)^\top$ (vermelho), $\boldsymbol{\psi}_2 = (\lambda = 1{,}0;\ \alpha = 0{,}2;\ \gamma = -1{,}0)^\top$ (verde), $\boldsymbol{\psi}_3 = (\lambda = 1{,}0;\ \alpha = 0{,}0;\ \gamma = 0{,}4)^\top$ (azul) .

A Figura 2 mostra algumas formas das funções de sobrevivência e da fdp da distribuição GTDL.

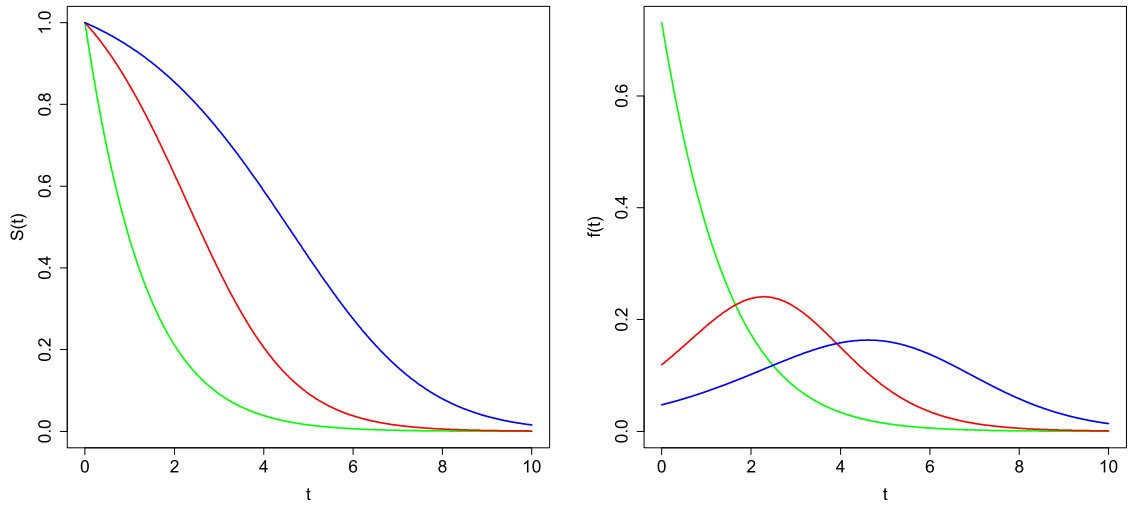


Figura 2: Gráficos das funções de sobrevivência do modelo GTDL (à esquerda) e da fdp (à direita): $\boldsymbol{\psi}_1 = (\lambda = 1{,}0;\ \alpha = 0{,}25;\ \gamma = 1{,}0)^\top$ (verde), $\boldsymbol{\psi}_2 = (\lambda = 1{,}0;\ \alpha = 0{,}75;\ \gamma = -2{,}0)^\top$ (vermelho), $\boldsymbol{\psi}_3 = (\lambda = 1{,}0;\ \alpha = 0{,}5;\ \gamma = -3{,}0)^\top$ (azul).

Ao analisar a Figura 2, observa-se que a função densidade de probabilidade exibe dife-

rentes formas com padrões variados tanto em termos de assimetria quanto de curtose. A análise demonstra que os parâmetros λ, α e γ influenciam significativamente o formato da função densidade e da função de sobrevivência. Particularmente, o parâmetro α mostra-se determinante nesta modelagem: valores maiores de α estão associados a tempos de falha mais concentrados em períodos iniciais, resultando em uma fdp com pico mais acentuado e uma função de sobrevivência com decaimento mais rápido. Por outro lado, valores menores de α correspondem a tempos de falha mais distribuídos ao longo do tempo, produzindo uma função densidade mais achatada e uma função de sobrevivência com declínio mais gradual.

O conceito de distribuição de probabilidade defeituosa é essencial na modelagem de fração de cura, permitindo capturar cenários em que parte da população não está sujeita ao evento de interesse. Nesse contexto, o modelo GTDL defeituoso amplia sua aplicabilidade ao incorporar a possibilidade de fração de cura, tornando-se uma ferramenta robusta para análises onde a proporcionalidade dos riscos não se mantém e a presença de indivíduos curados ou imunes é relevante.

3.2 O Modelo GTDL defeituoso

O modelo GTDL defeituoso corresponde ao modelo GTDL somente para valores de $\alpha < 0$. Essa característica é essencial na modelagem de fração de cura, especialmente em situações nas quais uma parte da população não experimenta o evento de interesse, como falha, morte ou recorrência de uma doença. Nesses casos, a função de sobrevivência não converge para zero quando t tende ao infinito, mas sim para um valor positivo ρ, que representa a fração de cura. Essa propriedade permite capturar a presença de indivíduos imunes ou curados, que nunca experimentarão o evento, diferenciando-se das distribuições tradicionais, que assumem que todos os indivíduos estão sujeitos ao evento.

Esse modelo oferece uma abordagem robusta para a análise de sobrevivência com fração de cura, sem a necessidade de modificar a função de sobrevivência, pois incorpora explicitamente a existência de indivíduos curados, considerando-os imunes ao evento. Essa abordagem é vantajosa por não requerer parâmetros adicionais, como ocorre em modelos tradicionais de fração de cura. Por exemplo, no trabalho de Cancho *et al.* (2011), uma abordagem Bayesiana é utilizada para analisar dados de sobrevivência, o que envolve a especificação de distribuições a priori para os parâmetros do modelo. Embora essa metodologia seja poderosa, ela pode ser mais complexa e computacionalmente intensiva em comparação com o modelo proposto por Mackenzie (1996).

Recentemente, modelos de fração de cura têm ganhado destaque na análise de sobrevivência, particularmente em cenários onde a presença de sobreviventes de longo prazo (indivíduos curados) e heterogeneidades populacionais é evidente. Um exemplo notável é o trabalho de Ramires *et al.* (2020), que propôs um modelo bimodal flexível capaz de capturar diferentes estruturas de regressão e incorporar explicitamente a fração de cura. Esse modelo é particularmente útil em situações onde a população pode ser dividida em subgrupos com comportamentos distintos, como pacientes que respondem a um tratamento versus aqueles que não respondem.

Essa abordagem é especialmente relevante em estudos médicos e de confiabilidade, onde a identificação de subpopulações com diferentes riscos é crucial para a tomada de decisões.

Para $\alpha < 0$, o modelo GTDL defeituoso apresenta uma fração de cura. Ou seja, quando $\alpha < 0$, tem-se:

$$\lim_{t \to +\infty} S(t, \psi) = \rho < 1.$$

A expressão da fração de cura ρ é central nessa abordagem, como discutido em Ibrahim *et al.*

(2013):

$$\begin{aligned} \rho &= \lim_{t\to+\infty} S(t,\psi) = \lim_{t\to+\infty} \left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}} \\ &= (1+\exp(\gamma))^{\lambda/\alpha} \in (0,1). \end{aligned} \tag{3.4}$$

Segundo Mackenzie (1996), para dados censurados, esta é uma propriedade desejável, pois, em muitos estudos, os tempos de falha mais longos tendem a ser censurados, e, portanto, a função de sobrevivência não decai até 0.

Para ilustrar o cálculo da fração de cura modelo GTDL defeituoso, considere os seguintes valores dos parâmetros $\lambda = 0,15$, $\alpha = -0,3$, e $\gamma = 3,00$. Com esses valores, obtém-se uma fração de cura $\rho = 0,22$, o que indica que 22% da população não experimentará o evento de interesse, mesmo após longos períodos de tempo. Esse exemplo destaca o comportamento típico do modelo quando $\alpha < 0$, refletindo a presença de indivíduos curados.

A Figura 3 apresenta a função de sobrevivência associada a esse cenário, evidenciando como ela se estabiliza em $\rho = 0{,}22$, o que representa a proporção de indivíduos que nunca falharão.

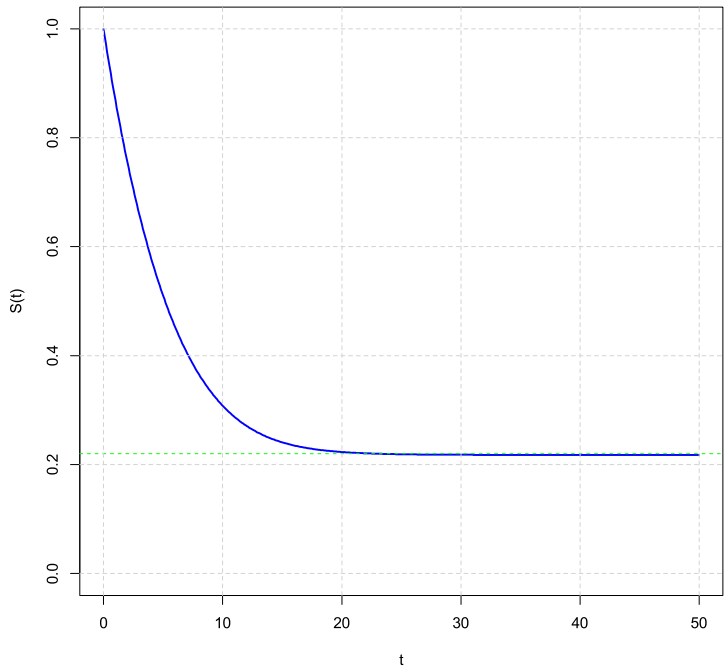


Figura 3: Gráfico da função de sobrevivência do modelo GTDL defeituoso com fração de cura $\rho = 0,22$.

Para $\alpha < 0$, como demonstrado em (3.4), pode-se deduzir que a função de distribuição acumulada do modelo GTDL defeituoso converge assintoticamente para

$$\lim_{t\to\infty} F(t) = 1 - (1+\exp(\gamma))^{\frac{\lambda}{\alpha}}.$$

Conforme detalhado no Capítulo 2, Seção 2.6, esta propriedade implica que o modelo possui uma fração de cura intrínseca ρ. A fdp correspondente é imprópria, característica fundamental dos modelos de fração de cura. Segundo Mackenzie (1996), é possível obter uma versão própria da função densidade do modelo GTDL defeituoso por meio da normalização. Esse processo ajusta a fdp de modo a garantir que a integral total seja igual a 1. É relevante ressaltar que a normalização da fdp no modelo GTDL defeituoso só se faz necessária em contextos onde a fração de curados não é de interesse. Quando o foco está na modelagem de uma fração de curados, a densidade imprópria é perfeitamente válida, uma vez que ρ é determinada diretamente pelo limite da função de sobrevivência quando $t \to \infty$. Isso resulta em uma fdp e a sua respectiva fda definida como:

$$f^*(t;\boldsymbol{\psi}) = \frac{\lambda\left(\frac{\exp(\alpha t+\gamma)}{1+\exp(\alpha t+\gamma)}\right)\left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}}{1-(1+\exp(\gamma))^{\frac{\lambda}{\alpha}}} \quad e$$

$$F^*(t;\boldsymbol{\psi}) = P(T \leq t) = 1 - P(T > t) = 1 - S(t,\boldsymbol{\psi}) = \frac{1-\left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}}{1-(1+\exp(\gamma))^{\frac{\lambda}{\alpha}}},$$

onde $\boldsymbol{\psi} = (\lambda, \alpha, \gamma)^T$ é o vetor de parâmetros, sendo $\lambda > 0$ um escalar, $\alpha < 0$, $\gamma \in \mathbb{R}$.

O comportamento monotônico da função de risco, já discutido para $\alpha \geq 0$, também se verifica quando $\alpha < 0$. Neste caso, a monotonicidade se expressa por um decréscimo da função, indicando que a taxa de falha diminui ao longo do tempo. A Figura 4 mostra a monoticidade da função de risco normalizado do modelo GTDL.

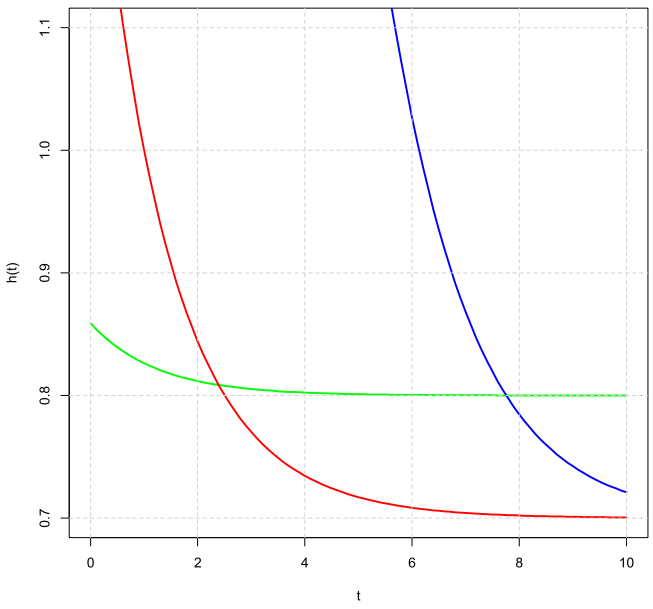


Figura 4: Gráfico normalizado da função de risco do modelo GTDL para diferentes vetores de parâmetros: $\boldsymbol{\theta}_1 = (\lambda = 18{,}0;\ \alpha = -0{,}8;\ \gamma = -5{,}0)^\top$ (verde), $\boldsymbol{\theta}_2 = (\lambda = 3{,}0;\ \alpha = -0{,}7;\ \gamma = 3{,}0)^\top$ (azul), $\boldsymbol{\theta}_3 = (\lambda = 9{,}0;\ \alpha = -0{,}7;\ \gamma = -2{,}0)^\top$ (vermelho).

A análise da Figura 4, com base nos valores dos parâmetros, revela diferentes comportamentos para a função de risco. As curvas apresentam um risco decrescente, indicando que o risco de falha diminui ao longo do tempo. Enquanto algumas curvas exibem um decaimento mais acentuado, com o risco diminuindo rapidamente, outras mostram um decaimento mais suave, refletindo uma redução gradual do risco ao longo do tempo.

Esses resultados ressaltam a flexibilidade do modelo GTDL em representar diferentes dinâmicas de risco e sobrevivência, o que é particularmente relevante em cenários onde o risco varia ao longo do tempo.

Segundo Mackenzie (1996), o modelo GTDL apresenta um comportamento próximo ao modelo de Gompertz. Essa aproximação entre os modelos pode ser demonstrada a partir da análise da função de risco do modelo GTDL, definida em (3.3). A aproximação ocorre sob a condição de que

$$\exp(\alpha t + \gamma) < 1,$$

o que permite considerar que o denominador pode ser aproximado por

$$1 + \exp(\alpha t + \gamma) \approx 1.$$

Substituindo essa aproximação na função (3.3), tem-se

$$\begin{aligned} h_{\text{GTDL}}(t) &\approx \lambda \left(\frac{\exp(\alpha t + \gamma)}{1} \right) \\ &= \lambda \exp(\alpha t + \gamma) \\ &= \lambda \exp(\gamma) \exp(\alpha t). \end{aligned}$$

Por sua vez, a função de risco do modelo de Gompertz é dada por

$$h_{\text{Gompertz}}(t) = a \exp(bt)$$

Igualando as duas expressões, obtém-se que

$$\begin{aligned} \lambda \exp(\gamma) \exp(\alpha t) &= a \exp(bt) \\ \Rightarrow a &= \lambda \exp(\gamma) \\ b &= \alpha \end{aligned}$$

Dessa forma, o coeficiente temporal b do modelo de Gompertz corresponde diretamente ao parâmetro α do modelo GTDL, enquanto o termo de escala a é dado por $\lambda \exp(\gamma)$. Vale destacar que Mackenzie (1996) não assume $\gamma = 0$. O termo $\exp(\gamma)$ é absorvido no parâmetro a do modelo Gompertz.

A Tabela 3.1 sintetiza as relações matemáticas entre os modelos GTDL e Gompertz,

Tabela 3.1: Comparação entre os modelos GTDL e Gompertz

Modelo	**Função de Risco**	**Condição**
GTDL Exato	$\lambda \left(\frac{\exp(\alpha t + \gamma)}{1 + \exp(\alpha t + \gamma)} \right)$	Geral
GTDL Aprox.	$\lambda \exp(\alpha t + \gamma)$	$\exp(\alpha t + \gamma) < 1$
Gompertz	$a \exp(bt)$	$a = \lambda \exp(\gamma)$, $b = \alpha$

Essa aproximação permite que o modelo GTDL capture dinâmicas de risco exponenciais semelhantes às da Gompertz, mantendo sua flexibilidade para incorporar covariáveis e frações de cura.

Por outro lado, o modelo de Gompertz apresenta limitações por assumir uma forma fixa para a função de risco, caracterizada por um crescimento exponencial. Dessa forma, o modelo GTDL se destaca por sua maior versatilidade na modelagem de dados de sobrevivência, alinhando-se às discussões da literatura sobre modelos de risco, conforme apresentado em Kalbfleisch e Prentice (2011).

3.3 O Modelo de regressão GTDL

A incorporação de covariáveis ao modelo GTDL é essencial para avaliar o impacto de fatores específicos sobre a sobrevivência. O modelo de regressão GTDL permite quantificar a influência dessas variáveis, proporcionando uma interpretação mais detalhada das características que afetam a duração do evento estudado. Segundo Mackenzie (1996), no modelo de regressão GTDL, o parâmetro γ da fdp (3.2), é substituído por $\boldsymbol{X}^\top\boldsymbol{\beta}$.

Seja T uma variável aleatória com distribuição GTDL, que representa o tempo até a ocorrência do evento de interesse. O modelo de regressão GTDL é caracterizado por apresentar as funções densidade de probabilidade, de sobrevivência e de risco, respectivamente, como:

$$f(t;\boldsymbol{\psi}) = \left(\lambda\frac{\exp\{\alpha t+\boldsymbol{X}^\top\boldsymbol{\beta}\}}{1+\exp\{\alpha t+\boldsymbol{X}^\top\boldsymbol{\beta}\}}\right)\times\left(\frac{1+\exp\{\alpha t+\boldsymbol{X}^\top\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}^\top\boldsymbol{\beta}\}}\right)^{-\lambda/\alpha},$$

$$S(t;\boldsymbol{\psi}) = \left(\frac{1+\exp\{\alpha t+\boldsymbol{X}^\top\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}^\top\boldsymbol{\beta}\}}\right)^{-\lambda/\alpha}$$

e

$$h(t;\boldsymbol{\psi}) = \lambda\frac{\exp\{\alpha t+\boldsymbol{X}^\top\boldsymbol{\beta}\}}{1+\exp\{\alpha t+\boldsymbol{X}^\top\boldsymbol{\beta}\}}.$$

A função de risco acumulada, conforme MacKenzie e Peng (2014), é dada por:

$$H(t;\boldsymbol{X}) = \int_0^t h(s\mid\boldsymbol{X})ds = \frac{\lambda}{\alpha}\log\left(\frac{1+\exp\{\alpha t+\boldsymbol{X}^\top\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}^\top\boldsymbol{\beta}\}}\right),$$

em que $\boldsymbol{\psi} = (\lambda,\alpha,\boldsymbol{\beta})^\top$ representa o vetor de parâmetros do modelo, $\lambda > 0$, o escalar $\alpha \in \mathbb{R}-\{0\}$ controla o efeito do tempo. O vetor $\boldsymbol{\beta} = (\beta_1,\beta_2,\ldots,\beta_p)^\top$ contém os coeficientes que medem a influência das p covariáveis no modelo, enquanto $\boldsymbol{X}^\top = (x_1,x_2,\ldots,x_p)$ é o vetor de covariáveis como por exemplo, idade, tratamento, entre outras. A expressão $\boldsymbol{X}^\top\boldsymbol{\beta}$ corresponde a uma combinação linear das covariáveis, representando o efeito das variáveis explicativas no modelo.

De acordo com Mackenzie (1996), o modelo de regressão GTDL não pode ser considerado como um modelo de riscos proporcionais ou como um modelo de vida acelerada. Essa distinção fica evidente ao se analisar a razão de riscos entre duas observações quaisquer, i e j, $i \neq j$, com diferentes valores de covariáveis $\boldsymbol{X}_i$ e $\boldsymbol{X}_j$, para $i,j = 1,\ldots,n$.

A razão de riscos entre essas duas observações é expressa por:

$$\frac{h(t|\boldsymbol{X}_i)}{h(t|\boldsymbol{X}_j)} = \frac{\lambda\dfrac{\exp(\alpha t + \boldsymbol{X}_i^\top\boldsymbol{\beta})}{1+\exp(\alpha t + \boldsymbol{X}_i^\top\boldsymbol{\beta})}}{\lambda\dfrac{\exp(\alpha t + \boldsymbol{X}_j^\top\boldsymbol{\beta})}{1+\exp(\alpha t + \boldsymbol{X}_j^\top\boldsymbol{\beta})}} = \frac{1+\exp(\alpha t + \boldsymbol{X}_j^\top\boldsymbol{\beta})}{1+\exp(\alpha t + \boldsymbol{X}_i^\top\boldsymbol{\beta})}\exp[(\boldsymbol{X}_i - \boldsymbol{X}_j)^\top\boldsymbol{\beta}].$$

A razão de riscos inclui um termo dependente do tempo, dado por:

$$\frac{1+\exp\{\alpha t + \boldsymbol{X}_j^\top\boldsymbol{\beta}\}}{1+\exp\{\alpha t + \boldsymbol{X}_i^\top\boldsymbol{\beta}\}}.$$

Esse termo evidencia que a razão entre os riscos muda ao longo do tempo, o que demonstra que o efeito do tempo não desaparece. Assim, os riscos não são proporcionais, uma vez que a relação entre os riscos não permanece constante para diferentes tempos t. Essa característica é fundamental para diferenciar o modelo GTDL de outros modelos, como o de riscos proporcionais de Cox ou de vida acelerada, nos quais a razão de riscos é invariável ao longo do tempo. A escolha entre os modelos depende do comportamento dos dados e das questões de pesquisa: o modelo de regressão Cox é uma escolha robusta para cenários com riscos proporcionais, enquanto o modelo de regressão GTDL é mais apropriado para dados com não proporcionalidade dos riscos ou presença de frações de cura.

Para explorar as propriedades estatísticas do modelo de regressão GTDL e obter estimativas para os parâmetros $\boldsymbol{\psi}$, é necessário maximizar a função de verossimilhança.

Dados (t_1, d_1), $(t_2, d_2), \ldots, (t_n, d_n)$ pares de observações de tamanho n de uma variável aleatória $T > 0$, caracterizada pelo modelo GTDL e d um indicador de censura. A função de verossimilhança para $T_1, T_2, \ldots, T_n$, independentes e identicamente distribuídas, considerando dados censurados é dada por:

$$L(\boldsymbol{\psi}) = \prod_{i=1}^{n}\left(\lambda\frac{\exp\{\alpha t_i + \boldsymbol{X}_i^\top\boldsymbol{\beta}\}}{1+\exp\{\alpha t_i + \boldsymbol{X}_i^\top\boldsymbol{\beta}\}}\right)^{d_i}\left(\frac{1+\exp\{\alpha t_i + \boldsymbol{X}_i^\top\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}_i^\top\boldsymbol{\beta}\}}\right)^{-\lambda/\alpha}.$$

O logaritmo da função de verossimilhança para o modelo de regressão GTDL é dado por:

$$\begin{aligned}\ell(\boldsymbol{\psi}) &= \sum_{i=1}^{n} d_i\left[\log(\lambda) + (\alpha t_i + \boldsymbol{X}_i^\top\boldsymbol{\beta}) - \log\left(1+\exp\{\alpha t_i + \boldsymbol{X}_i^\top\boldsymbol{\beta}\}\right)\right] \\ &- \frac{\lambda}{\alpha}\sum_{i=1}^{n}\left[\log\left(1+\exp\{\alpha t_i + \boldsymbol{X}_i^\top\boldsymbol{\beta}\}\right) - \log\left(1+\exp\{\boldsymbol{X}_i^\top\boldsymbol{\beta}\}\right)\right].\end{aligned}$$

Os estimadores de máxima verossimilhança $\hat{\boldsymbol{\psi}} = (\hat{\lambda}, \hat{\alpha}, \hat{\beta})$ são obtidos resolvendo o sistema de equações diferenciais parciais:

$$\frac{\partial\ell(\boldsymbol{\psi})}{\partial\lambda} = 0, \quad \frac{\partial\ell(\boldsymbol{\psi})}{\partial\alpha} = 0, \quad \frac{\partial\ell(\boldsymbol{\psi})}{\partial\beta} = 0.$$

Esse sistema de equações não lineares em relação aos parâmetros, não possuem soluções fechadas explícitas. Portanto, métodos numéricos são necessários para encontrar as estimativas. Entre os métodos mais utilizados estão o algoritmo de Newton-Raphson, o método de escore de Fisher e o método Quasi-Newton. Neste trabalho utiliza-se o método Quasi-Newton implementado na função `optim` da linguagem de programação R (R Core Team, 2024).

Capítulo 4

Modelo Logístico Generalizado Estendido Dependente do Tempo (GTDEL)

Neste capítulo, apresenta-se a distribuição GTDEL, abordando sua formulação matemática e propriedades, bem como suas principais funções. Em seguida, introduz-se a distribuição GTDEL defeituosa, que incorpora a possibilidade de fração de cura. Discute-se também o método de estimação dos parâmetros do modelo. Por fim, apresentam-se a função quantil e um estudo de simulação para avaliar o comportamento dos estimadores de máxima verossimilhança.

4.1 Formulação da distribuição GTDEL

Na literatura especializada, diversas técnicas têm sido propostas para a construção de famílias de distribuições generalizadas. Dentre essas, destaca-se a família exponencializada-F (Exp-F). Essa família é definida ao elevar a função de distribuição acumulada $F(t;\boldsymbol{\psi})$ a uma potência $\delta > 0$. Esse procedimento gera uma nova classe de distribuições de probabilidade, cuja fda é dada por

$$G(t;\boldsymbol{\theta}) = (F(t;\boldsymbol{\psi}))^{\delta}, \tag{4.1}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\psi};\delta)^{\top}$ representa o vetor de parâmetros da nova família de distribuições; $\boldsymbol{\psi} = (\lambda,\alpha,\gamma)^{\top}$ é o vetor de parâmetros da distribuição base; $G(t;\boldsymbol{\theta})$ corresponde à função de distribuição acumulada do novo modelo; e δ é o novo parâmetro introduzido.

Consequentemente, a fdp da nova distribuição é obtida diferenciando com respeito a t a Expressão (4.1):

$$g(t;\boldsymbol{\theta}) = \delta f(t;\boldsymbol{\psi})F(t;\boldsymbol{\psi})^{\delta-1},$$

em que, $f(t;\boldsymbol{\psi})$ é a fdp da distribuição base.

A função de sobrevivência da nova distribuição pode ser escrita da seguinte forma:

$$S(t;\boldsymbol{\theta}) = 1 - F(t;\boldsymbol{\psi})^{\delta}.$$

A função de risco correspondente é:

$$h(t;\boldsymbol{\theta}) = \frac{\delta f(t;\boldsymbol{\theta}) F(t;\boldsymbol{\theta})^{\delta-1}}{1 - F(t;\boldsymbol{\theta})^{\delta}}.$$

A vantagem dessa nova classe de distribuições é que a inclusão do parâmetro δ amplia a versatilidade do modelo para se ajustar aos dados observados de forma mais precisa, melhorando o desempenho do modelo em termos de ajuste e precisão e a capacidade de incorporar um número maior de submodelos da distribuição que foi generalizada. Essa abordagem tem sido explorada em trabalhos como o de Pascoa (2012).

Em geral, as distribuições generalizadas são mais flexíveis que a distribuição de origem. Além de englobarem a distribuição base como caso particular, elas oferecem maior potencial analítico para investigação das propriedades de cauda, conforme discutido em Gomes (2024).

A família em (4.1) vem sendo utilizada por muito autores para construir novas classes de distribuições que são extensões dos modelos usuais. Dentro dessa classe de distribuições, cita-se a Weibull Exponenciada, apresentada por Mudholkar *et al.* (1996), que é uma generalização da distribuição Weibull. Gupta e Kundu (2001), propuseram a distribuição Exponencial Generalizada que é uma generalização da distribuição Exponencial. Carrasco *et al.* (2008), utilizaram para estender a distribuição Weibull. Shawky e Bakoban (2008), discutiram a distribuição Gama Exponenciada que é uma generalização da distribuição gama. Cordeiro *et al.* (2011), propuseram a distribuição gama generalizada exponenciada com aplicação aos dados de tempo de vida. Afify e Abdellatif (2020), generalizaram a distribuição Burr XII usando o método de exponenciação e aplicaram em dois conjuntos de dados reais de sobrevivência, entre outras.

A distribuição GTDEL constitui uma generalização da distribuição GTDL, permitindo maior flexibilidade na modelagem da dependência temporal do risco. Esta extensão teórica é particularmente útil em estudos de sobrevivência onde a razão de risco não se mantém constante ao longo do tempo, permitindo um ajuste mais refinado às características dos dados.

4.2 O Modelo Probabilístico GTDEL

Para um escalar $\delta > 0$, a função de distribuição acumulada de probabilidade do novo modelo, chamado de Modelo Logístico Generalizado Estendido Dependente do Tempo (GTDEL) é obtida pela substituição da Equação (3.1) na Equação (4.1). Assim, a fda do novo modelo é definido por

$$G(t;\boldsymbol{\theta}) = \left[1 - \left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta},$$

e sua fdp, função de sobrevivência e risco são expressas, respectivamente, por:

$$g(t;\boldsymbol{\theta}) = \lambda\delta\left(\frac{\exp(\alpha t+\gamma)}{1+\exp(\alpha t+\gamma)}\right)\left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}\left[1 - \left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta-1} \quad (4.2)$$

$$S(t;\boldsymbol{\theta}) = 1 - \left[1 - \left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta},$$

$$h\left(t;\boldsymbol{\theta}\right) = \frac{\lambda\delta\left(\dfrac{\exp\left(\alpha t+\gamma\right)}{1+\exp\left(\alpha t+\gamma\right)}\right)\left(\dfrac{1+\exp\left(\alpha t+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}\left[1-\left(\dfrac{1+\exp\left(\alpha t+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta-1}}{1-\left[1-\left(\dfrac{1+\exp\left(\alpha t+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta}},$$

onde $\boldsymbol{\theta} = (\boldsymbol{\psi}; \delta)^\top$ é o vetor de parâmetros da nova distribuição.

Na Figura 1 são exibidas algumas formas possíveis da fdp da distribuição GTDEL para diferentes valores paramétricos.

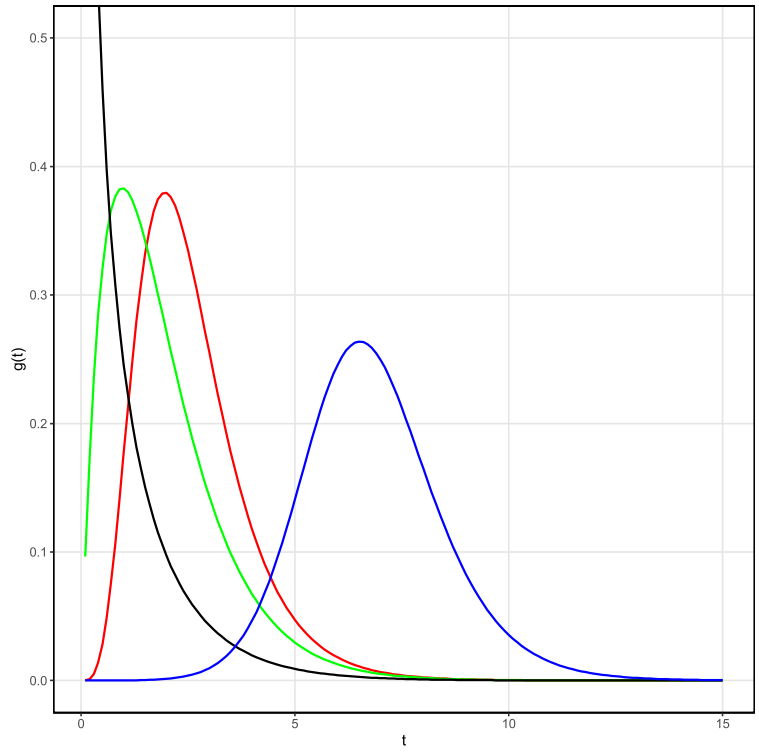


Figura 1: Formas da função densidade de probabilidade da distribuição GTDEL para diferentes vetores de parâmetros: $\boldsymbol{\theta}_1 = (\lambda = 1{,}0;\ \alpha = 0{,}8;\ \gamma = 1{,}0;\ \delta = 5{,}0)^\top$ (vermelho), $\boldsymbol{\theta}_2 = (\lambda = 1{,}0;\ \alpha = 0{,}2;\ \gamma = 1{,}0;\ \delta = 2{,}0)^\top$ (verde), $\boldsymbol{\theta}_3 = (\lambda = 1{,}0;\ \alpha = 0{,}7;\ \gamma = -4{,}0;\ \delta = 2{,}9)^\top$ (azul) e $\boldsymbol{\theta}_4 = (\lambda = 1{,}0;\ \alpha = 0{,}05;\ \gamma = 1{,}0;\ \delta = 0{,}5)^\top$ (preto).

A análise do gráfico apresentado na Figura 1, indica que a distribuição GTDEL permite diferentes comportamentos da fdp, dependendo dos valores dos parâmetros. Para examinar o efeito desses parâmetros na forma da fdp, a Figura 1 exibe três cenários distintos: cauda leve (preto), moderada (verde) e pesada (vermelho).

Observa-se que, para valores menores de α e δ, a fdp apresenta uma cauda mais leve, indicando uma diminuição mais rápida da probabilidade de eventos em valores mais altos. À medida que os valores de α e δ aumentam, a distribuição passa a apresentar caudas mais pesadas, o que significa que a probabilidade de eventos em valores mais altos diminui de forma mais lenta. Esse comportamento é uma característica importante da distribuição GTDEL, permitindo que se ajuste a diferentes tipos de dados.

Na Figura 2, são apresentadas diferentes formas da função de risco para distintos valores paramétricos da distribuição GTDEL.

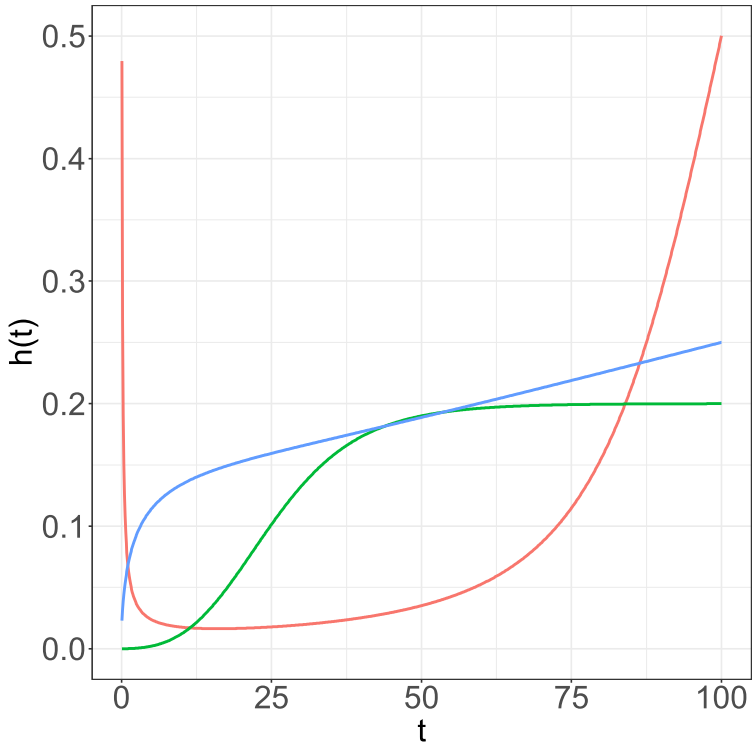


Figura 2: Formas da função de risco da distribuição GTDEL para diferentes vetores de parâmetros: $\boldsymbol{\theta}_1 = (\lambda = 1{,}0;\ \alpha = 0{,}09;\ \gamma = -9{,}0;\ \delta = 0{,}1)^\top$ (vermelho), $\boldsymbol{\theta}_2 = (\lambda = 0{,}2;\ \alpha = 0{,}1;\ \gamma = -2{,}0;\ \delta = 2{,}9)^\top$ (verde) e $\boldsymbol{\theta}_3 = (\lambda = 0{,}5;\ \alpha = 0{,}01;\ \gamma = -1{,}0;\ \delta = 1{,}5)^\top$ (azul).

Observa-se, na Figura 2, a presença de distintas formas da função de risco: uma curva em formato aproximado de banheira (vermelho), o comportamento sigmoidal (verde) e crescente (azul). A análise dos gráficos evidencia que a introdução do parâmetro de forma δ trouxe um aprimoramento significativo nas possíveis formas de risco da distribuição.

A forma de banheira é amplamente utilizada em Análise de Sobrevivência, pois descreve três estágios fundamentais do ciclo de vida de um sistema ou indivíduo:

1. Fase inicial: Alto risco de falha (mortalidade precoce),
2. Fase intermediária: Risco relativamente constante e reduzido (fase de vida útil),
3. Fase final: Risco crescente devido ao envelhecimento ou desgaste (mortalidade tardia).

Além disso, a distribuição GTDEL é capaz de representar padrões de risco com comportamento sigmoidal. Essa característica destaca-se pela capacidade do modelo em capturar variações suaves e não lineares ao longo do tempo, tornando-o adequado para diversas aplicações em modelagem de sobrevivência, especialmente em cenários onde o risco varia de forma gradual e contínua.

Com isso, a distribuição GTDEL se apresenta como uma ferramenta robusta e versátil para estudos de confiabilidade e sobrevivência, adaptando-se a diferentes cenários e padrões de risco. A possibilidade de incorporar múltiplas formas – como a banheira, a sigmoidal e crescente – amplia seu potencial de aplicação na análise de dados complexos em diversas áreas.

A Figura 3 mostra um exemplo da função de distribuição acumulada da distribuição GTDEL.

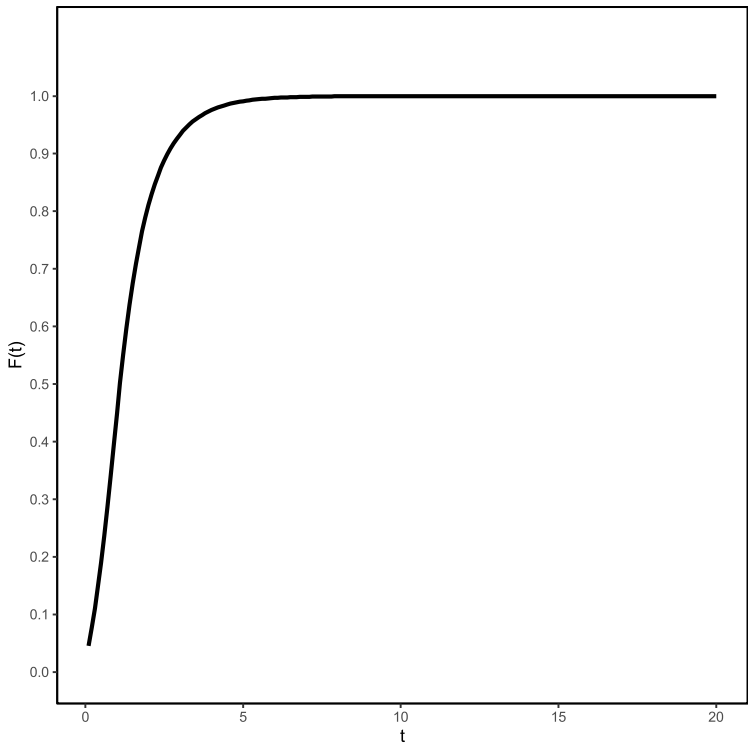


Figura 3: Função de distribuição acumulada da distribuição GTDEL para os seguintes valores paramétricos: $\boldsymbol{\theta}_1 = (\lambda = 1{,}0;\ \alpha = 3{,}9;\ \gamma = -3{,}0;\ \delta = 0{,}6)^\top$.

A distribuição GTDEL possui a propriedade de englobar submodelos conhecidos como casos particulares quando se impõem restrições específicas a seus parâmetros. Pode-se observar que:

1. Se $\lambda = \delta = 1$ o modelo GTDEL se reduz à família logística tempo-dependente (TDL) com fdp dada por:

$$g\left(t;\boldsymbol{\theta}\right) \quad = \quad \left(\frac{\exp\left(\alpha t+\gamma\right)}{1+\exp\left(\alpha t+\gamma\right)}\right)\left(\frac{1+\exp\left(\alpha t+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{1}{\alpha}}.$$

2. Se $\delta = 1$ obtém-se o modelo inicial GTDL com fdp:

$$g\left(t;\boldsymbol{\theta}\right) \quad = \quad \lambda\left(\frac{\exp\left(\alpha t+\gamma\right)}{1+\exp\left(\alpha t+\gamma\right)}\right)\left(\frac{1+\exp\left(\alpha t+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}.$$

Apesar da variabilidade introduzida pelo parâmetro de forma δ, a distribuição GTDEL

herda algumas limitações da distribuição GTDL. Especificamente, quando $\alpha < 0$, a fdp pode tornar-se imprópria, o que compromete a interpretação probabilística do modelo em determinadas situações.

Na próxima seção, será discutido esse problema no contexto da distribuição GTDEL e a estratégia de normalização adotada para garantir que a função densidade seja própria, assegurando sua aplicabilidade em modelagem de sobrevivência.

4.3 A distribuição GTDEL defeituosa

A distribuição GTDEL defeituosa corresponde à distribuição GTDEL apenas para valores de α menores que zero, incorporando assim a possibilidade de uma fração de cura. Isso permite modelar cenários em que uma proporção da população estudada não irá experimentar o evento de interesse, mesmo após longos períodos de acompanhamento. Essa característica se dá quando a função de sobrevivência não atinge zero, mas se estabiliza em um valor ρ, que representa a proporção de indivíduos imunes ao evento, isto é,

$$\lim_{t \to +\infty} S\left(t; \boldsymbol{\theta}\right) = \rho > 0, \tag{4.3}$$

em que $\rho \in (0, 1)$.

A Figura 4 ilustra a fda da distribuição defeituosa GTDEL.

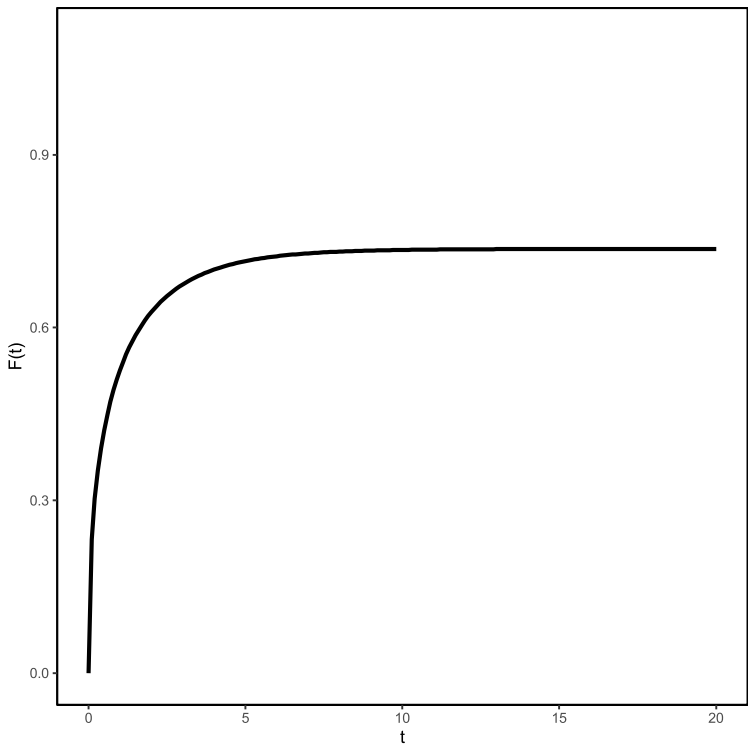


Figura 4: Função de distribuição acumulada da distribuição defeituosa GTDEL para $\boldsymbol{\theta}_1 = (\lambda = 1{,}0;\ \alpha = -0{,}5;\ \gamma = 1{,}0;\ \delta = 0{,}4)^\top$.

Conforme se vê na Figura 4, o comportamento da função de distribuição acumulada destaca a presença de indivíduos imunes ao evento, assumindo um valor assintoticamente igual a $\rho < 1$. Esse comportamento é fundamental para modelar corretamente dados de sobrevivência com cura, pois permite diferenciar os indivíduos suscetíveis dos verdadeiramente curados.

A distribuição GTDEL para $\alpha < 0$, é uma distribuição defeituosa, pois permite a incorporação de uma fração de cura ρ, que é calculada como,

$$\begin{aligned}
\rho &= \lim_{t\to+\infty} S(t,\boldsymbol{\theta}) \\
&= \lim_{t\to+\infty} 1 - \left[1 - \left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta} \\
&= \lim_{t\to+\infty} 1 - \left[1 - \left(\frac{1}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}} (1+\exp(\alpha t+\gamma))^{-\frac{\lambda}{\alpha}}\right]^{\delta} \\
&= 1 - \left[1 - \left(\frac{1}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}} \cdot \lim_{t\to+\infty} (1+\exp(\alpha t+\gamma))^{-\frac{\lambda}{\alpha}}\right]^{\delta}.
\end{aligned}$$

Como,

$$\lim_{t\to+\infty} (1+\exp(\alpha t+\gamma))^{\frac{-\lambda}{\alpha}} = 1,$$

então,

$$\rho = 1 - [1 - (1+\exp(\gamma))^{\frac{\lambda}{\alpha}}]^{\delta} \in (0,1). \tag{4.4}$$

É importante ressaltar que a condição $\alpha < 0$ é necessária para que o limite da função de sobrevivência (4.3) convirja para a a proporção de curados ρ e esteja no intervalo $(0,1)$.

Quando $\alpha > 0$, não existe fração de cura, pois o limite da função de sobrevivência (4.3) converge para 0. Essa condição, invalida a interpretação de ρ como uma proporção de curados, pois uma fração de curados igual a zero implicaria que todos os indivíduos estão sujeitos ao evento de interesse, o que contradiz a premissa básica de um modelo de fração de cura.

Portanto, a distribuição GTDEL defeituosa só permite o cálculo de ρ quando $\alpha < 0$, garantindo que a proporção de curados esteja no intervalo $(0,1)$ e seja interpretável no contexto de modelos de sobrevivência com fração de cura.

Para ilustrar a existência da fração de cura, considere o caso particular em que $\lambda = 1,2$, $\alpha = -0,5$, $\gamma = -1,5$ e $\delta = 0,5$. Substituindo esses valores paramétricos em (4.4), temos $\rho = 0,38$, indicando que 38% da população não falhará, mesmo ao longo de períodos muito longos. Na Figura 5 é apresentado o gráfico da função de sobrevivência da distribuição GTDEL defeituosa, com essa fração de cura.

Na Figura 6, são apresentadas diferentes formas da função de sobrevivência do modelo GTDEL defeituoso, considerando os parâmetros fixos $\lambda = 0,2$, $\alpha = -1,5$, $\gamma = -1$ e distintos valores de δ. Observa-se que a fração de cura ρ é diretamente influenciada pelo parâmetro δ: à medida que $\delta \to +\infty$, a fração de cura se aproxima de 1, enquanto, quando $\delta \to 0$, ρ tende a 0. Esse comportamento evidencia que valores mais elevados de δ estão associados a uma maior proporção de indivíduos curados, ou seja, menos indivíduos experimentam o evento de interesse.

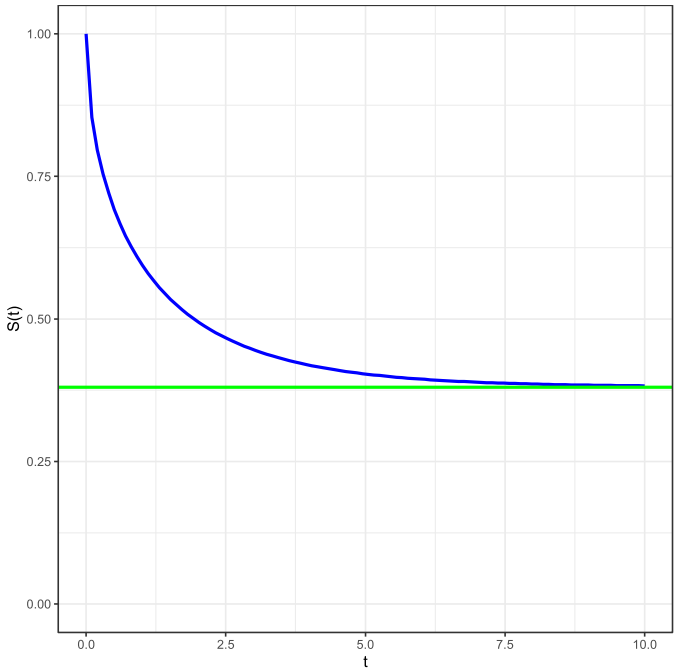


Figura 5: Gráfico da função de sobrevivência do modelo GTDEL defeituoso com fração de cura $\rho = 38\%$.

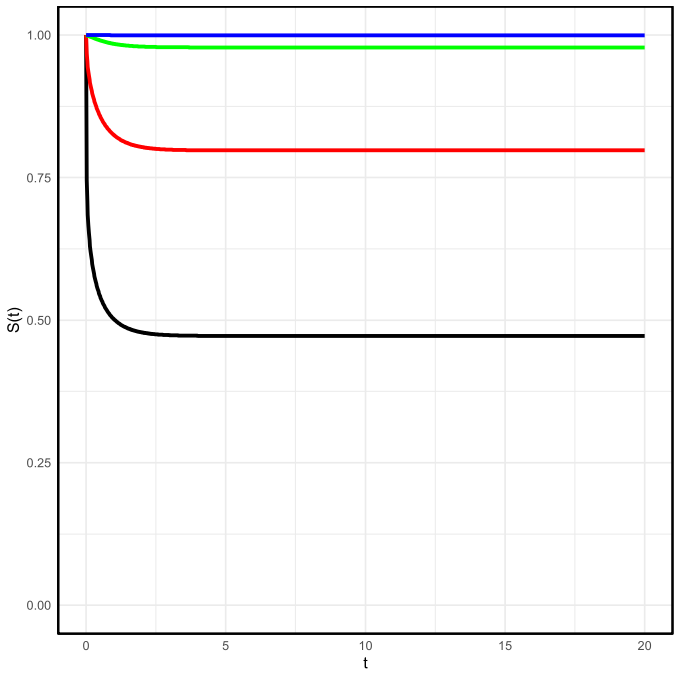


Figura 6: Gráfico da função de sobrevivência do modelo GTDEL defeituoso para diferentes valores de $\delta = 0,2$ (preto), $\delta = 0,5$ (vermelho), $\delta = 1,2$(verde) e $\delta = 2,5$ (azul).

A distribuição GTDEL é uma generalização da distribuição GTDL e, por isso, preserva suas principais características estruturais. Para $\alpha < 0$ assim como na distribuição GTDL, a fda da GTDEL não é completamente definida no intervalo $(0, 1)$, resultando em uma função densidade de probabilidade imprópria. Para garantir que a fdp seja própria, é necessário um

processo de normalização semelhante ao realizado no modelo GTDL (Mackenzie, 1996). Esse ajuste é feito dividindo a função densidade original pelo limite da função de sobrevivência quando $t \to \infty$, assegurando que a integral total da fdp seja igual a 1.

É importante destacar que a necessidade de normalização da fdp ocorre especificamente quando se trabalha com $\alpha < 0$, pois, nessa condição, a função densidade resultante torna-se imprópria. A normalização é exigida apenas quando se deseja utilizar a distribuição como um modelo de sobrevivência convencional, ou seja, em contextos nos quais todos os indivíduos eventualmente experimentam o evento de interesse. No entanto, quando o objetivo da modelagem é considerar a existência de uma fração de curados na população, a utilização de uma densidade imprópria é admissível. Nesses casos, a fração de curados, representada por ρ, é calculada diretamente como o limite da função de sobrevivência quando o tempo tende ao infinito, o que dispensa a normalização da fdp.

A seguir, são apresentadas as funções de distribuição acumulada e densidade de probabilidade do modelo GTDEL normalizado para $\alpha < 0$.

$$G^*(t;\boldsymbol{\theta}) = \left[\frac{1-\left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}}{1-(1+\exp(\gamma))^{\frac{\lambda}{\alpha}}}\right]^{\delta}.$$

$$g^*(t;\boldsymbol{\theta}) = \frac{\lambda\delta\left(\frac{\exp(\alpha t+\gamma)}{1+\exp(\alpha t+\gamma)}\right)\left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}\left[1-\left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta-1}}{\left[1-(1+\exp(\gamma))^{\frac{\lambda}{\alpha}}\right]^{\delta}},$$

onde $\boldsymbol{\theta} = (\lambda, \alpha, \gamma, \delta)^\top$, $\lambda > 0$, $\alpha < 0$, $\gamma \in \mathbb{R}$ e $\delta > 0$.

4.4 Estimação dos parâmetros

A metodologia utilizada para a estimação dos parâmetros do modelo GTDEL é o método de máxima verossimilhança discutido na Seção 2.4. O método da máxima verossimilhança, torna-se adequado para a estimação, pois em grandes amostras possui propriedades desejáveis além dos demais métodos.

4.4.1 Estimação na ausência de censura

Suponha $t_1,\ t_2, \ldots,\ t_n$ uma amostra aleatória de tamanho n do modelo GTDEL. O logaritmo da função de verossimilhança é dado por:

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= n\log\lambda + n\log\delta + \alpha\sum_{i=1}^{n} t_i + n\gamma - \left(\frac{\lambda}{\alpha}+1\right)\sum_{i=1}^{n}\log\left[1+\exp\left(\alpha t_i+\gamma\right)\right] \\ &+ \frac{n\lambda}{\alpha}\log\left[1+\exp\left(\gamma\right)\right] + (\delta-1)\sum_{i=1}^{n}\log\left[1-\left(\frac{1+\exp\left(\alpha\, t_i+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}\right].\end{aligned} \tag{4.5}$$

Derivando a função de log-verossimilhança $\ell(\boldsymbol{\theta})$ com relação aos parâmetros λ, α, γ e δ, obtêm-se os elementos do vetor escore $U(\boldsymbol{\theta})$. O vetor escore é definido como o gradiente da log-verossimilhança em relação aos parâmetros, isto é:

$$U(\boldsymbol{\theta}) = \frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \left(\frac{\partial\ell(\boldsymbol{\theta})}{\partial\lambda}, \frac{\partial\ell(\boldsymbol{\theta})}{\partial\alpha}, \frac{\partial\ell(\boldsymbol{\theta})}{\partial\gamma}, \frac{\partial\ell(\boldsymbol{\theta})}{\partial\delta}\right)^{\top},$$

em que

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\theta})}{\partial\lambda} =& \frac{n}{\lambda} - \frac{1}{\alpha}\sum_{i=1}^{n}\log\left[1+\exp\left(\alpha t_i+\gamma\right)\right] + \frac{n}{\alpha}\log\left(1+\exp\left(\gamma\right)\right) \\ &+ \frac{(\delta-1)}{\alpha}\sum_{i=1}^{n}\frac{\left(\frac{1+\exp\left(\alpha t_i+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}\log\left(\frac{1+\exp\left(\alpha t_i+\gamma\right)}{1+\exp\left(\gamma\right)}\right)}{1-\left(\frac{1+\exp\left(\alpha t_i+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}}\end{aligned} \tag{4.6}$$

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\theta})}{\partial\alpha} =& \sum_{i=1}^{n} t_i - \frac{\lambda n}{\alpha^2}\log\left[1+\exp\left(\gamma\right)\right] - \left(\frac{\lambda}{\alpha}+1\right)\sum_{i=1}^{n}\frac{t_i\exp\left(\alpha t_i+\gamma\right)}{1+\exp\left(\alpha t_i+\gamma\right)} \\ &+\frac{\lambda\left(1-\delta\right)}{\alpha}\sum_{i=1}^{n}\frac{\left(\frac{1+\exp\left(\alpha t_i+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}}{1-\left(\frac{1+\exp\left(\alpha t_i+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}}\left[\frac{1}{\alpha}\log\left(\frac{1+\exp\left(\alpha t_i+\gamma\right)}{1+\exp\left(\gamma\right)}\right) - \frac{t_i\exp\left(\alpha t_i+\gamma\right)}{1+\exp\left(\alpha t_i+\gamma\right)}\right] \\ &+\frac{\lambda}{\alpha^2}\sum_{i=1}^{n}\log\left[1+\exp\left(\alpha t_i+\gamma\right)\right]\end{aligned} \tag{4.7}$$

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\theta})}{\partial\gamma} &= \frac{\lambda\left(\delta-1\right)}{\alpha}\sum_{i=1}^{n}\frac{\left[\exp\left(\alpha t_i+\gamma\right)-\exp\left(+\gamma\right)\right]\left(\frac{1+\exp\left(\alpha t_i+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}-1}}{\left[1+\exp\left(\gamma\right)\right]^2\left[1-\left(\frac{1+\exp\left(\alpha t_i+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}\right]} \\ &+ n - \left(\frac{\lambda}{\alpha}+1\right)\sum_{i=1}^{n}\frac{\exp\left(\alpha t_i+\gamma\right)}{1+\exp\left(\alpha t_i+\gamma\right)} + \frac{n\lambda\exp\left(+\gamma\right)}{\alpha\left[1+\exp\left(+\gamma\right)\right]}\end{aligned} \tag{4.8}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \delta} = \frac{n}{\delta} + \sum_{i=1}^{n} \log\left[1 - \left(\frac{1+\exp(\alpha\, t_i + \gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}\right]. \tag{4.9}$$

O estimador de máxima verossimilhança (EMV), denotado por $\widehat{\boldsymbol{\theta}}$, é obtido igualando as Equações 4.6−4.9 a zero e resolvendo-as simultaneamente. É conveniente utilizar algoritmos de otimização para equações não lineares. Particularmente, neste trabalho é utilizado o método Quasi-Newton desenvolvido por Broyden-Fletcher-Goldfarb-Shanno (método BFGS), que está implementado na função optim(·, method = "BFGS") do *software* (R Core Team, 2024).

Os erros-padrão e intervalos de confiança assintóticos, são encontrados assumindo que a distribuição do estimador de máxima verossimilhança é aproximadamente normal multivariado. Quando $n \to \infty$, a distribuição assintótica do EMV é dada por:

$$\begin{pmatrix} \widehat{\lambda} \\ \widehat{\alpha} \\ \widehat{\gamma} \\ \widehat{\delta} \end{pmatrix} \approx N\left[\begin{pmatrix} \lambda \\ \alpha \\ \gamma \\ \delta \end{pmatrix}, \begin{pmatrix} \widehat{V}_{11} & \widehat{V}_{12} & \widehat{V}_{13} & \widehat{V}_{14} \\ \widehat{V}_{21} & \widehat{V}_{22} & \widehat{V}_{23} & \widehat{V}_{24} \\ \widehat{V}_{31} & \widehat{V}_{32} & \widehat{V}_{33} & \widehat{V}_{34} \\ \widehat{V}_{41} & \widehat{V}_{42} & \widehat{V}_{43} & \widehat{V}_{44} \end{pmatrix}\right],$$

sendo que $N(\cdot\,,\cdot)$ representa a distribuição normal multivariada, $\widehat{V}_{ij} = V_{ij}|_{\Theta=\widehat{\Theta}}$ e

$$\begin{pmatrix} \widehat{V}_{11} & \widehat{V}_{12} & \widehat{V}_{13} & \widehat{V}_{14} \\ \widehat{V}_{21} & \widehat{V}_{22} & \widehat{V}_{23} & \widehat{V}_{24} \\ \widehat{V}_{31} & \widehat{V}_{32} & \widehat{V}_{33} & \widehat{V}_{34} \\ \widehat{V}_{41} & \widehat{V}_{42} & \widehat{V}_{43} & \widehat{V}_{44} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{pmatrix}^{-1}$$

é a matriz de variância e covariância, cujos elementos são dados por $A_{rs} = \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial\theta_r\, \partial\theta_s}$, para $r, s = 1, 2, 3, 4$.

Intervalos de confiança bilaterais com $100(1-\breve{\alpha})\%$ de confiança para λ, α, γ e δ são, respectivamente, dados por:

$$\widehat{\lambda} \pm z_{\breve{\alpha}/2}\sqrt{\widehat{V}_{11}}, \widehat{\alpha} \pm z_{\breve{\alpha}/2}\sqrt{\widehat{V}_{22}}, \widehat{\gamma} \pm z_{\breve{\alpha}/2}\sqrt{\widehat{V}_{33}}, \widehat{\delta} \pm z_{\breve{\alpha}/2}\sqrt{\widehat{V}_{44}},$$

sendo que $z_{\breve{\alpha}}$ é o percentil superior de ordem σ da distribuição normal padrão.

4.4.2 Estimação na presença de censura

Considere uma amostra aleatória (t_i, d_i), onde $i = 1, \ldots, n$, com $t_i = min(T_i, C_i)$ e

$$d_i = \begin{cases} 1, & \text{se } T_i \leq C_i \text{ (observação não censurada)}, \\ 0, & \text{se } T_i > C_i \text{ (observação censurada)}. \end{cases}$$

Onde, T_i representa o tempo de falha e C_i o tempo de censura, assumidos como independentes. A função do logaritmo de verossimilhança para o modelo GTDEL na presença de dados censurados é expressa por:

$$
\begin{aligned}
\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} d_i \log\delta + \alpha\sum_{i=1}^{n} d_i t_i + \sum_{i=1}^{n} d_i\gamma - \left(\frac{\lambda}{\alpha}+1\right)\sum_{i=1}^{n} d_i \log\left[1+\exp\left(\alpha t_i+\gamma\right)\right] \\
&+ \frac{\lambda}{\alpha}\sum_{i=1}^{n} d_i \log\left[1+\exp\left(\gamma\right)\right] + (\delta-1)\sum_{i=1}^{n} d_i \log\left[1-\left(\frac{1+\exp\left(\alpha\, t_i+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}\right] \\
&+ (1-d_i)\sum_{i=1}^{n} \log\left\{1-\left[1-\left(\frac{1+\exp\left(\alpha t+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta}\right\} + \sum_{i=1}^{n} d_i \log\lambda.
\end{aligned}
$$

Neste cenário, a função de verossimilhança incorpora tanto as observações completas (não censuradas) quanto as censuradas. Os estimadores de máxima verossimilhança são obtidos maximizando a função $\ell(\boldsymbol{\theta})$. Para isso, resolve-se o sistema de equações formado pelas derivadas parciais de $\ell(\boldsymbol{\theta})$ em relação aos parâmetros $\boldsymbol{\theta} = (\lambda, \alpha, \gamma, \delta)^{\top}$, ou seja,

$$
U(\boldsymbol{\theta}) = \frac{\partial \log \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}.
$$

4.5 Função quantil e geração de amostra aleatória da distribuição GTDEL

O quantil de uma distribuição é um conceito fundamental na análise de dados, utilizado para dividir e entender a distribuição de um conjunto de observações em partes iguais ou percentis específicos. Em análise de sobrevivência, os quantis são particularmente úteis para identificar pontos críticos na distribuição do tempo de sobrevivência, como a mediana, que divide a população em duas partes iguais.

Para $\alpha > 0$, o quantil de ordem p, denotado por t_p da distribuição GTDEL, é definido como a solução real da equação:

$$
G\left(t_p\right) = p,
$$

em que $G(t_p)$ representa a função de distribuição acumulada da distribuição GTDEL. Daí,

$$
\left[1-\left(\frac{1+\exp\left(\alpha t_p+\gamma\right)}{1+\exp\left(\gamma\right)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta} = p.
$$

A solução da equação acima fornece o quantil t_p da distribuição GTDEL. Isolando t_p, obtém-se a seguinte expressão:

$$
t_p = \frac{\ln[(1-p^{\frac{1}{\delta}})^{-\frac{\alpha}{\lambda}}(1+\exp(\gamma))-1]-\gamma}{\alpha}.
$$

Em particular, a mediana da distribuição $t_{0,5}$ é calculada da seguinte forma:

$$t_{0,5} = \frac{\ln[(1-(0,5)^{\frac{1}{\delta}})^{\frac{-\alpha}{\lambda}}(1+\exp(\gamma))-1]-\gamma}{\alpha}.$$

Por outro lado, o quantil de ordem p, da distribuição GTDEL para $\alpha < 0$, na ausência de censura é dado por:

$$G^*(t_p) = p,$$

onde $G^*(t_p)$ representa a função de distribuição acumulada defeituosa normalizada da distribuição GTDEL. Assim, temos:

$$\left[\frac{1-\left(\frac{1+\exp(\alpha t_p+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}}{1-(1+\exp(\gamma))^{\frac{\lambda}{\alpha}}}\right]^{\delta} = p.$$

Isolando a variável t_p na equação acima, obtém-se a seguinte expressão para o quantil:

$$t_p = \frac{\ln[\{1-[1-(1+\exp(\gamma))^{\frac{\lambda}{\alpha}}]p^{\frac{1}{\delta}}\}^{-\frac{\alpha}{\lambda}}(1+\exp(\gamma))-1]-\gamma}{\alpha}.$$

Algoritmo para geração de amostras aleatórias

Diversos métodos são propostos na literatura para gerar dados aleatórios, incluindo técnicas amplamente utilizadas, como o método de Aceitação- Rejeição, o algoritmo de Metropolis-Hastings e o método de Box-Muller (Casella e Berger, 2024). Neste trabalho, optou-se pelo método da Transformação Inversa, descrito em Magalhães (2006), para gerar amostras aleatórias da distribuição GTDEL.

O método da transformação inversa é um método clássico para gerar amostras de uma distribuição contínua, a partir de sua fda. A ideia central consiste em explorar o fato de que, se $U \sim \text{Uniform}(0,1)$, então a variável aleatória $X = F^{-1}(U)$ segue a distribuição com fda $F(x)$. Dessa forma, a geração de observações passa pela obtenção de valores de U e pela aplicação da inversa da fda. Esse método é particularmente eficaz quando a fda e sua inversa podem ser expressas de forma analítica, como no caso da distribuição GTDEL. No entanto, para distribuições cuja inversa não é obtida de maneira fechada, a aplicação prática pode ser inviável, sendo necessário recorrer a métodos alternativos.

Para $\alpha > 0$, o procedimento consiste em:

1. Gerar uma variável uniforme $u \sim U(0,1)$;
2. Aplicar a inversa da fda para obter a amostra: $t = G^{-1}(u)$.

A fda inversa do modelo GTDEL é obtida a partir da expressão:

$$u = \left[1 - \left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta},$$

então,

$$t = \frac{\ln[(1-u^{\frac{1}{\delta}})^{-\frac{\alpha}{\lambda}}(1+\exp(\gamma))-1]-\gamma}{\alpha}. \tag{4.10}$$

Por outro lado, no caso em que $\alpha < 0$, a geração dos dados pode ser realizada com ou sem a consideração explícita da censura, incorporando diretamente a presença de indivíduos curados.

Na ausência de censura, as amostras foram geradas utilizando a função de distribuição acumulada normalizada $G^*(t)$ para gerar amostras por meio do método da inversa, ou seja,

$$t = G^{*-1}(p),$$

em que $p \sim U(0,1)$ é uma variável aleatória uniforme. Para esse caso, u é solução da seguinte equação:

$$u = \left[\frac{1-\left(\frac{1+\exp(\alpha t+\gamma)}{1+\exp(\gamma)}\right)^{-\frac{\lambda}{\alpha}}}{1-(1+\exp(\gamma))^{\frac{\lambda}{\alpha}}}\right]^{\delta}.$$

Isolando t, obtém-se:

$$t = \frac{\ln[\{1-[1-(1+\exp(\gamma))^{\frac{\lambda}{\alpha}}]u^{\frac{1}{\delta}}\}^{-\frac{\alpha}{\lambda}}(1+\exp(\gamma))-1]-\gamma}{\alpha}. \tag{4.11}$$

Por sua vez, a geração de dados considerando explicitamente a presença de uma fração de curados foi realizada com base na Equação (4.4), que expressa a fração de cura do modelo GTDEL.

A Figura 7 apresenta os resultados de uma simulação da distribuição GTDEL sem considerar a presença de censura, abrangendo duas situações distintas: à esquerda, com $\alpha > 0$, e à direita, com $\alpha < 0$. Em ambos os casos, a função densidade de probabilidade da distribuição GTDELé sobreposta ao histograma das amostras simuladas, o que permite avaliar o ajuste do modelo aos dados gerados.

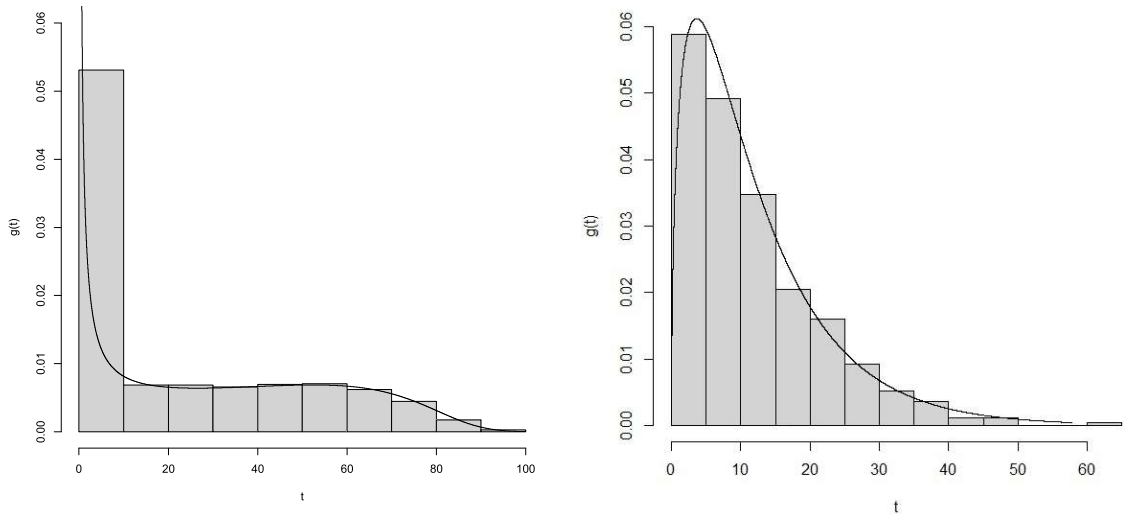


Figura 7: Gráficos da simulação do modelo GTDEL: à esquerda, para $\alpha > 0$ com parâmetros $\lambda = 1{,}0$, $\alpha = 0{,}09$, $\gamma = -9{,}0$ e $\delta = 0{,}1$; à direita, para $\alpha < 0$ com $\lambda = 0{,}45$, $\alpha = -0{,}1$, $\gamma = -3{,}0$ e $\delta = 1{,}5$.

Observa-se na Figura 7 que, para $\alpha > 0$, a função densidade do modelo GTDEL assume uma forma assimétrica, com cauda mais longa à direita, refletindo maior concentração de eventos de falha no início do período observado. A curva ajustada segue bem o comportamento do histograma, indicando que o modelo reproduz adequadamente a distribuição dos dados simulados.

No cenário com $\alpha < 0$, a função densidade também se ajusta satisfatoriamente ao histograma, respeitando as características de assimetria e curtose presentes na amostra gerada. O bom desempenho visual reforça a flexibilidade da distribuição GTDEL em diferentes configurações paramétricas.

4.6 Estudo de Simulação

Nesta secão, apresenta-se um estudo de simulação com o objetivo de conhecer o comportamento assintótico dos estimadores de máxima verossimilhança da distribuição GTDEL. Em cada amostra, os parâmetros do modelo foram estimados via maximização direta do logaritmo da função de verossimilhança, utilizando o algoritmo L-BFGS-B (`optim`) do software R (R Core Team, 2024). O estudo visa avaliar a consistência e a precisão dos estimadores para diferentes tamanhos amostrais, além de verificar o desempenho do modelo em cenários controlados.

O método de simulação Monte Carlo, conforme descrito por Robert e Casella (2010), é empregado para gerar amostras da distribuição GTDEL. Essa abordagem é particularmente útil em situações onde as condições teóricas são complexas, permitindo a análise numérica do comportamento dos estimadores em cenários controlados.

A geração dos dados foi realizada com base na Equação (4.10) para o caso em que $\alpha > 0$. Para $\alpha < 0$, o procedimento varia conforme a presença ou ausência de censura: na ausência de censura, utiliza-se a Equação (4.11); já na presença de censura e fração de cura, aplica-se a estrutura definida pela Equação (4.4).

Foram considerados três tamanhos amostrais: $n = 500$, 1000 e 2000. Para cada um deles, foram realizadas $R = 1000$ réplicas de Monte Carlo, sob quatro diferentes cenários.

Adotando-se o vetor de parâmetros $\boldsymbol{\theta} = (\lambda, \alpha, \gamma, \delta)^\top$, são calculadas, para cada parâmetro $\boldsymbol{\theta}_j$ $(j = 1, 2, 3, 4)$, as seguintes métricas:

1. Média das Estimativas:

$$\bar{\theta}_j = \frac{1}{R}\sum_{i=1}^{R}\hat{\theta}_j^{(i)},$$

2. Viés:

$$\text{Viés}(\theta_j) = \bar{\theta}_j - \theta_j.$$

3. Raiz do Erro Quadrático Médio (REQM):

$$\text{REQM}(\theta_j) = \sqrt{\frac{1}{R}\sum_{i=1}^{R}(\hat{\theta}_j^{(i)} - \theta_j)^2},$$

   mede a precisão dos estimadores.

4. Taxa de Cobertura (TC):

$$\text{TC}(\theta_j) = \frac{\#\{\theta_j \in IC[\theta_j; 1-\breve{\alpha}]\}}{R},$$

que representa a porcentagem de vezes que o intervalo de confiança de 95% , contém o valor verdadeiro do parâmetro, onde $IC[\theta_j; 1-\breve{\alpha}]$ é o intervalo de confiança para θ_j com coeficiente nominal $1-\breve{\alpha}$.

As Tabelas 4.1 e 4.2 apresentam os resultados do estudo de simulação para os cenários 1 e 2, nos quais $\alpha > 0$ e $\alpha < 0$, respectivamente, desconsiderando a presença de censura. Para cada tamanho de amostra, são reportadas as médias das estimativas obtidas por máxima verossimilhança, os respectivos viés, raízes do erro quadrático médio e taxas de cobertura.

Para complementar a análise numérica dos estimadores obtidos por máxima verossimilhança, apresentados nas tabelas para os cenários 1, 2, 3 e 4, utilizaram-se boxplots das estimativas dos parâmetros do modelo GTDEL. Os boxplots são representações gráficas que sintetizam a distribuição de um conjunto de dados, destacando a mediana, os quartis, possíveis outliers e a dispersão geral. No contexto desta simulação, cada boxplot mostra a distribuição das estimativas obtidas em 1.000 réplicas para um dado parâmetro e tamanho amostral.

As linhas tracejadas vermelhas indicam os valores verdadeiros dos parâmetros, servindo como referência visual para avaliação do viés e da precisão das estimativas. Essa abordagem gráfica permite observar o comportamento dos estimadores em termos de concentração, viés e variação, além de facilitar comparações entre diferentes tamanhos amostrais em cada cenário.

Cenário 1 ($\alpha > 0$):

Consideramos neste cenário os seguintes valores verdadeiros para os parâmetros $\boldsymbol{\theta} = (0{,}2; 0{,}4; -3{,}0; 1{,}1)$.

A Tabela 4.1 apresenta os resultados para $\alpha > 0$, em que os dados são gerados sem a presença de censura.

Tabela 4.1: Média, viés, REQM e TC das estimativas dos parâmetros da distribuição GTDEL para o cenário 1, com $\alpha > 0$.

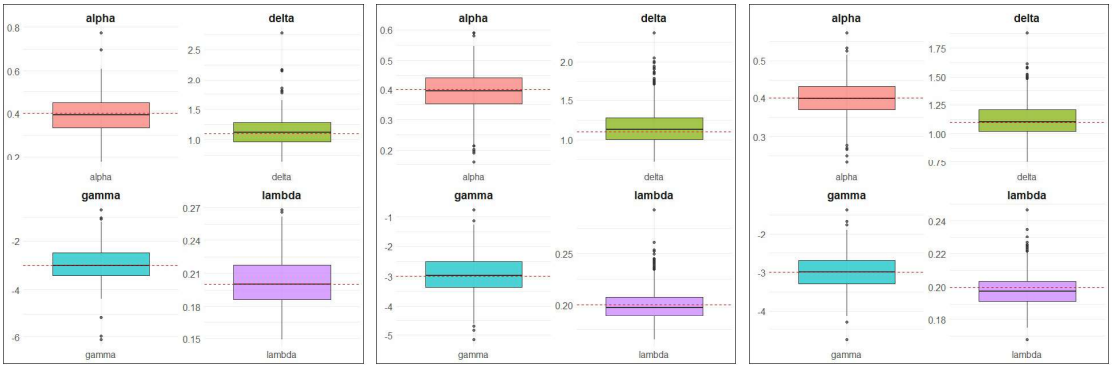
		Parâmetros			
n	Medida	λ	α	γ	δ
500	Média	$0,206$	$0,391$	$-2,915$	$1,203$
	Viés	$0,006$	$-0,009$	$0,085$	$0,103$
	REQM	$0,068$	$0,095$	$0,871$	$0,350$
	TC	$95,3$	$94,8$	$95,8$	$95,7$
1000	Média	$0,199$	$0,397$	$-2,959$	$1,152$
	Viés	$-0,001$	$-0,003$	$0,041$	$0,052$
	REQM	$0,014$	$0,063$	$0,633$	$0,226$
	TC	$94,9$	$95,2$	$94,5$	$95,6$
2000	Média	$0,198$	$0,400$	$-2,995$	$1,120$
	Viés	$-0,002$	$0,000$	$0,005$	$0,020$
	REQM	$0,009$	$0,046$	$0,446$	$0,145$
	TC	$95,0$	$94,9$	$94,9$	$94,6$

Na Tabela 4.1 observa-se que a qualidade das estimativas melhora à medida que o tamanho amostral aumenta. As médias das estimativas se aproximam dos valores dos parâmetros, o que evidencia a consistência dos estimadores.

Além disso, o viés e a raiz do erro quadrático médio diminuem com o aumento de n, indicando maior precisão e acurácia. Essa tendência é particularmente notável a partir de $n = 1000$, quando os valores de viés tornam-se bastante reduzidos e as REQMs apresentam os menores valores observados.

Em relação à taxa de cobertura dos intervalos de confiança de 95%, os resultados são satisfatórios em todos os cenários. Mesmo para tamanhos amostrais menores, os valores permanecem próximos ao nível nominal de 95%.

Na Figura 8 são apresentados os resultados para os casos em que $\alpha > 0$ (cenário 1).



(a) (b) (c)

Figura 8: Boxplots das estimativas dos parâmetros para diferentes tamanhos amostrais: (a) $n = 500$, (b) $n = 1000$ e (c) $n = 2000$ para o cenário 1.

Observa-se que, à medida que o tamanho amostral aumenta, os boxplots tornam-se mais concentrados em torno dos valores verdadeiros, com menor dispersão e redução no número de outliers. Esse comportamento confirma visualmente a consistência dos estimadores, já evidenciada pelos valores médios e pelas REQMs da Tabela 4.1.

Cenário 2 ($\alpha < 0$):

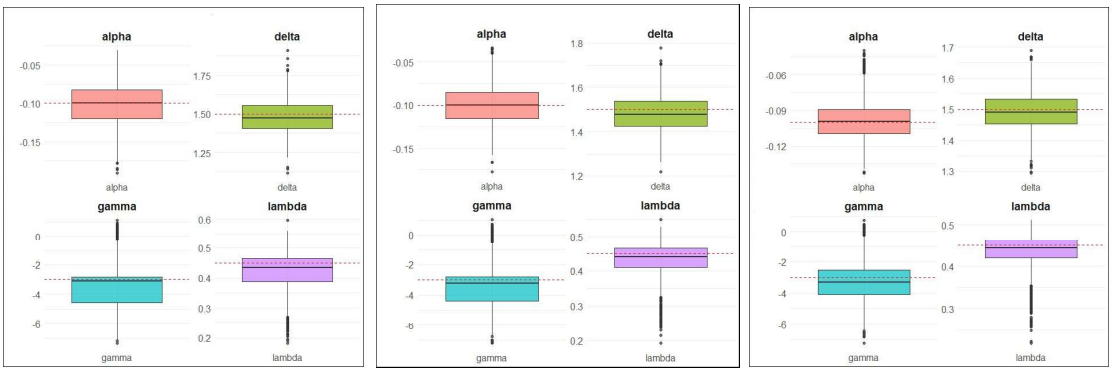
Neste cenário, considerou-se vetor de parâmetros $\boldsymbol{\theta} = (0{,}45; -0{,}1; -3{,}0; 1{,}5)$. A Tabela 4.2 apresenta os resultados do estudo de simulação para o cenário em que $\alpha < 0$.

Tabela 4.2: Média, viés, REQM e TC das estimativas dos parâmetros da distribuição GTDEL para o cenário 2, $\alpha < 0$.

		Parâmetros			
n	Medida	λ	α	γ	δ
	Média	0,418	−0,100	−3,389	1,482
	Viés	−0,032	−0,000	−0,389	−0,018
500	REQM	0,077	0,028	1,742	0,114
	TC	94,1	85,7	92,3	94,6
	Média	0,425	−0,099	−3,291	1,483
	Viés	−0,025	0,001	−0,291	−0,017
1000	REQM	0,063	0,023	1,631	0,083
	TC	93,7	87,2	90,6	92,6
	Média	0,433	−0,098	−3,175	1,492
	Viés	−0,017	0,002	−0,175	−0,008
2000	REQM	0,048	0,018	1,383	0,059
	TC	95,7	88,9	93,0	94,7

Como esperado, os resultados na Tabela 4.2, mostra um desempenho bom à medida que o tamanho amostral aumenta. As médias das estimativas aproximam-se dos valores verdadeiros com o crescimento de n, e os vieses tendem a diminuir. As REQMs também seguem essa tendência de redução. A taxa de cobertura dos intervalos de confiança são, em geral, satisfatórias, mesmo para tamanhos amostrais menores. Os resultados evidenciam que o método de estimação baseado na máxima verossimilhança fornece estimativas consistentes e com boas propriedades inferenciais.

Na Figura 9 são apresentados os resultados para o caso em que $\alpha < 0$ (cenário 2).



(a) (b) (c)

Figura 9: Boxplots das estimativas dos parâmetros para diferentes tamanhos amostrais: (a) $n = 500$, (b) $n = 1000$ e (c) $n = 2000$ para o cenário 2.

A análise gráfica apresentada na Figura 9 complementa os resultados numéricos da Tabela 4.2, ao evidenciar visualmente a concentração das estimativas em torno dos valores reais à medida que o tamanho amostral aumenta. Observa-se uma redução progressiva na dispersão das estimativas e no número de outliers, sinalizando maior precisão.

Destaca-se a variabilidade relativamente maior nas estimativas do parâmetro γ, especialmente para amostras menores, o que é compatível com os maiores REQMs observados na tabela. Por outro lado, o parâmetro α, apresenta estimativas bem concentradas e viés praticamente nulo, confirmando a robustez do estimador de máxima verossimilhança nesse contexto.

As Tabelas 4.3 e 4.4 apresentam os resultados do estudo de simulação para os cenários 3 e 4, correspondentes aos casos em que $\alpha > 0$ e $\alpha < 0$, respectivamente. Para o cenário com $\alpha > 0$, foram considerados três níveis distintos de censura: 10%, 30% e 50%. Já no cenário com $\alpha < 0$, a análise foi conduzida sob diferentes proporções de fração de cura, determinadas a partir de combinações específicas dos parâmetros do vetor $\boldsymbol{\theta} = (\lambda, \alpha, \gamma, \delta)^\top$.

Três configurações de parâmetros foram avaliadas no caso com $\alpha < 0$: a primeira, com $\boldsymbol{\theta} = (0{,}45,\ -0{,}01,\ -3{,}0,\ 1{,}50)$, resultou em uma fração de cura de aproximadamente 16%; a segunda, com $\boldsymbol{\theta} = (0{,}45,\ -0{,}02,\ -3{,}00,\ 1{,}50)$, correspondeu a 45%; e a terceira, com $\boldsymbol{\theta} = (0{,}45,\ -0{,}03,\ -3{,}00,\ 1{,}50)$, a 62%. As amostras foram geradas com base no algoritmo proposto por Rocha *et al.* (2016), que simula dados a partir de um modelo defeituoso.

Para cada combinação de cenário e tamanho amostral, são apresentados: as médias das estimativas obtidas por máxima verossimilhança, os respectivos vieses, as raízes do erro quadrático médio (REQM) e as taxas de cobertura dos intervalos de confiança de 95%.

Complementando a análise numérica, as Figuras 10 (caso $\alpha > 0$) e 11 (caso $\alpha < 0$) exibem os boxplots das estimativas obtidas, também estratificados pelos três níveis de censura ou frações de cura consideradas em cada situação.

Cenário 3 ($\alpha > 0$), na presença de censura:

Algoritmo de Geração de dados com Censura

1. Gera-se o tempo de falha t_1.

2. Gera-se o tempo de censura $t_2 \sim \text{Exp}(c)$, em que $c > 0$ controla a proporção de censura e $E[T_c] = 1/c$.

3. Define-se o tempo observado como $t = \min\{t_1, t_2\}$.

4. A variável indicadora de censura é definida como $d = 1$ se $t = t_1$ (evento observado) e $d = 0$ se $t = t_2$ (evento censurado).

A Tabela 4.3 a seguir apresenta os resultados do estudo de simulação para este cenário, considerando (10%, 30% e 50%) de censura. De modo geral, os resultados indicam que os estimadores do modelo GTDEL mantêm viés reduzido e média próxima aos valores reais dos parâmetros, sugerindo que o estimador de máxima verossimilhança é aproximadamente não viesado mesmo sob censura. Observa-se que os REQMs diminuem com o aumento do tamanho amostral, o que evidencia a consistência dos estimadores.

Como esperado, o aumento da proporção de censura provoca elevação dos REQMs, especialmente para os menores tamanhos amostrais, sendo o parâmetro γ o mais afetado por essa

variação. Ainda assim, o desempenho global dos estimadores permanece satisfatório. As taxas de cobertura dos intervalos de confiança de 95% se mantêm próximas ao nível nominal na maioria dos cenários, com pequenas quedas nas situações mais extremas (menor n e maior censura).

Vale destacar que, para amostras com tamanho superior a 1000, os estimadores apresentaram excelente desempenho, mesmo com 50% de censura, revelando REQMs baixos, viés próximo de zero e TCs próximas de 95%. Tais resultados confirmam a robustez do método de máxima verossimilhança aplicado ao modelo GTDEL em contextos com censura.

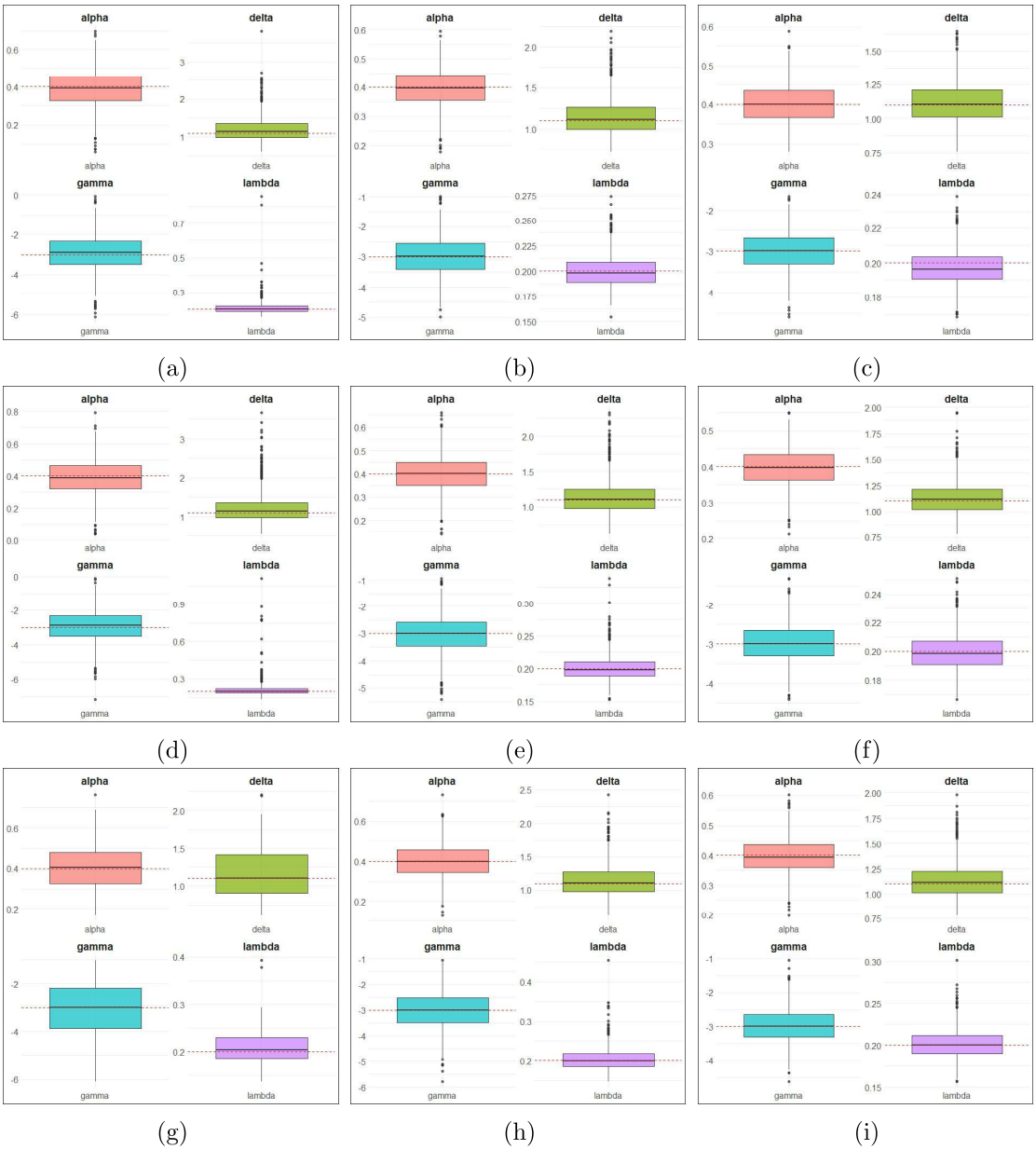


Figura 10: Boxplots das estimativas dos parâmetros do modelo GTDEL no cenário 3 ($\alpha > 0$), considerando três tamanhos amostrais (n = 500, 1000 e 2000) sob diferentes proporções de censura: 10% em (a)–(c), 30% em (d)–(f) e 50% em (g)–(i).

Os boxplots apresentados na Figura 10 permitem visualizar o comportamento das estimativas dos parâmetros do modelo GTDEL sob diferentes proporções de censura, no cenário em que $\alpha > 0$. As linhas tracejadas indicam os valores reais dos parâmetros utilizados na simulação,

Tabela 4.3: Média, viés, REQM e TC das estimativas dos parâmetros do modelo GTDEL para o cenário 3.

n	Medida	Censura											
		10%				30%				50%			
		λ	α	γ	δ	λ	α	γ	δ	λ	α	γ	δ
500	Média	0,201	0,390	-2,929	1,175	0,212	0,388	-2,929	1,219	0,228	0,389	-2,996	1,211
	Viés	0,001	-0,010	0,071	0,104	0,012	-0,012	0,071	0,119	0,028	-0,011	0,004	0,111
	REQM	0,036	0,098	0,885	0,336	0,064	0,111	0,955	0,411	0,192	0,126	1,004	0,394
	TC	96,8	94,9	95,1	95,4	96,9	94,4	94,3	94,5	96,7	93,2	95,7	95,5
1000	Média	0,201	0,390	-2,899	1,175	0,201	0,402	-3,017	1,140	0,204	0,399	-3,001	1,149
	Viés	0,001	-0,010	0,101	0,075	0,001	0,002	-0,017	0,040	0,004	-0,001	-0,001	0,049
	REQM	0,016	0,069	0,654	0,249	0,020	0,078	0,717	0,248	0,028	0,085	0,715	0,248
	TC	95,6	94,2	93,7	96,2	95,5	93,7	92,6	93,3	97,6	95,0	94,8	96,2
2000	Média	0,197	0,402	-3,007	1,117	0,200	0,398	-2,970	1,129	0,202	0,397	-2,973	1,135
	Viés	-0,003	0,002	-0,007	0,017	-0,000	-0,002	0,030	0,029	0,002	-0,003	0,027	0,035
	REQM	0,010	0,048	0,453	0,145	0,013	0,053	0,475	0,158	0,018	0,060	0,522	0,177
	TC	94,3	95,3	95,6	95,0	96,0	94,3	95,3	95,4	96,0	95,1	94,1	95,4

facilitando a identificação de viés e dispersão.

Observa-se que, mesmo na presença de censura, os boxplots mantêm a concentração das estimativas em torno dos valores verdadeiros, especialmente para tamanhos amostrais maiores. Esse comportamento é coerente com os baixos valores de viés apresentados na Tabela 4.3, indicando que os estimadores permanecem aproximadamente não-viesados em todos os níveis de censura.

Conforme esperado, os efeitos da censura se manifestam com mais intensidade nas amostras menores, onde se nota maior dispersão e presença de outliers, principalmente nas estimativas do parâmetro γ, o mais sensível à censura. Essa observação é respaldada pelos valores mais elevados de REQM para γ nos cenários com censura de 30% e 50%.

Cenário 4 ($\alpha < 0$), na presença de fração de cura:

Algoritmo de Geração de Dados do Modelo Defeituoso

Conforme descrito por Rocha *et al.* (2016), o algoritmo para geração de dados a partir de um modelo defeituoso é o seguinte:

1. Defina os valores dos parâmetros do modelo e a fração de cura desejada ρ.

2. Gere $M_i \sim \text{Bernoulli}(1-\rho)$.

3. Se $M_i = 0$, defina $t_i' = \infty$ (indivíduo curado). Caso contrário, tome t_i' como a raiz da equação $G(t) = u$, onde $u \sim \text{Uniforme}(0, 1-\rho)$.

4. Gere $u_i' \sim \text{Uniforme}(0, \max\{t_i\})$, considerando apenas os t_i finitos.

5. Calcule $t_i = \min(t_i', u_i')$ e defina $d_i = 1$ se $t_i < u_i'$, ou $d_i = 0$ caso contrário.

A Tabela 4.4 a seguir apresenta os resultados do estudo de simulação para o cenário em que $\alpha < 0$, considerando três diferentes configurações de parâmetros associadas a distintas proporções de cura: aproximadamente 16%, 45% e 62%. As estimativas dos parâmetros do modelo GTDEL mostram médias próximas dos valores reais utilizados na simulação em cada cenário, o que indica que o estimador de máxima verossimilhança é não viesado ou apresenta viés desprezível, especialmente para tamanhos amostrais maiores. Além disso, os valores do viés e da raiz do erro quadrático médio (REQM) diminuem com o aumento do tamanho amostral, evidenciando o ganho de precisão do estimador com n.

Observa-se também que, embora as configurações dos parâmetros variem entre os três cenários, as diferenças entre as estimativas se mantêm pequenas, mesmo com o aumento significativo da fração de cura. No entanto, nota-se que os valores do REQM tendem a ser maiores nos cenários com maior proporção de cura, o que pode ser atribuído ao fato de que, nos cenários com maior proporção de cura, uma parcela significativa dos indivíduos nunca experimenta o evento de interesse. Isso reduz a quantidade de informação disponível sobre os tempos de falha, dificultando a estimação precisa dos parâmetros relacionados à parte suscetível da população.

De modo geral, os resultados indicam que o estimador baseado na verossimilhança para o modelo GTDEL apresenta bom desempenho, com médias próximas dos valores verdadeiros, baixo viés e REQM decrescente com o aumento de n.

Tabela 4.4: Média, viés e REQM das estimativas dos parâmetros do modelo GTDEL para o cenário 4.

n	Medida	Censura											
		16%				45%				62%			
		λ	α	γ	δ	λ	α	γ	δ	λ	α	γ	δ
500	Média	0,459	-0,011	-2,995	1,517	0,455	-0,021	-2.992	1,513	0,493	-0,032	-2,959	1,550
	Viés	0,009	-0,001	0,005	0,017	0,005	-0,001	-0,008	0,013	-0,043	-0,002	0,041	0,050
	REQM	0,030	0,001	0,013	0,093	0,066	0,002	0,124	0,161	0,221	0,006	0,647	0,243
1000	Média	0,454	-0,010	-2,997	1,503	0,455	-0,020	-2,996	1,510	0,461	-0,031	-2,975	1,501
	Viés	0,004	-0,001	0,003	0,003	0,005	-0,001	0,004	0,010	-0,011	-0,001	-0,025	0,001
	REQM	0,016	0,001	0,007	0,014	0,043	0,001	0,029	0,114	0,158	0,003	0,546	0,146
2000	Média	0,451	-0,010	-2,999	1,501	0,452	-0,020	-3,00	1,504	0,453	-0,030	-2,974	1,505
	Viés	0,001	-0,001	0,001	0,001	0,002	-0,001	-0,001	0,004	-0,003	-0,001	-0,025	-0,005
	REQM	0,013	0,002	0,001	0,006	0,034	0,001	0,027	0,068	0,114	0,001	0,235	0,094

A Figura 11 apresenta os boxplots das estimativas dos parâmetros do modelo GTDEL para o cenário 4.

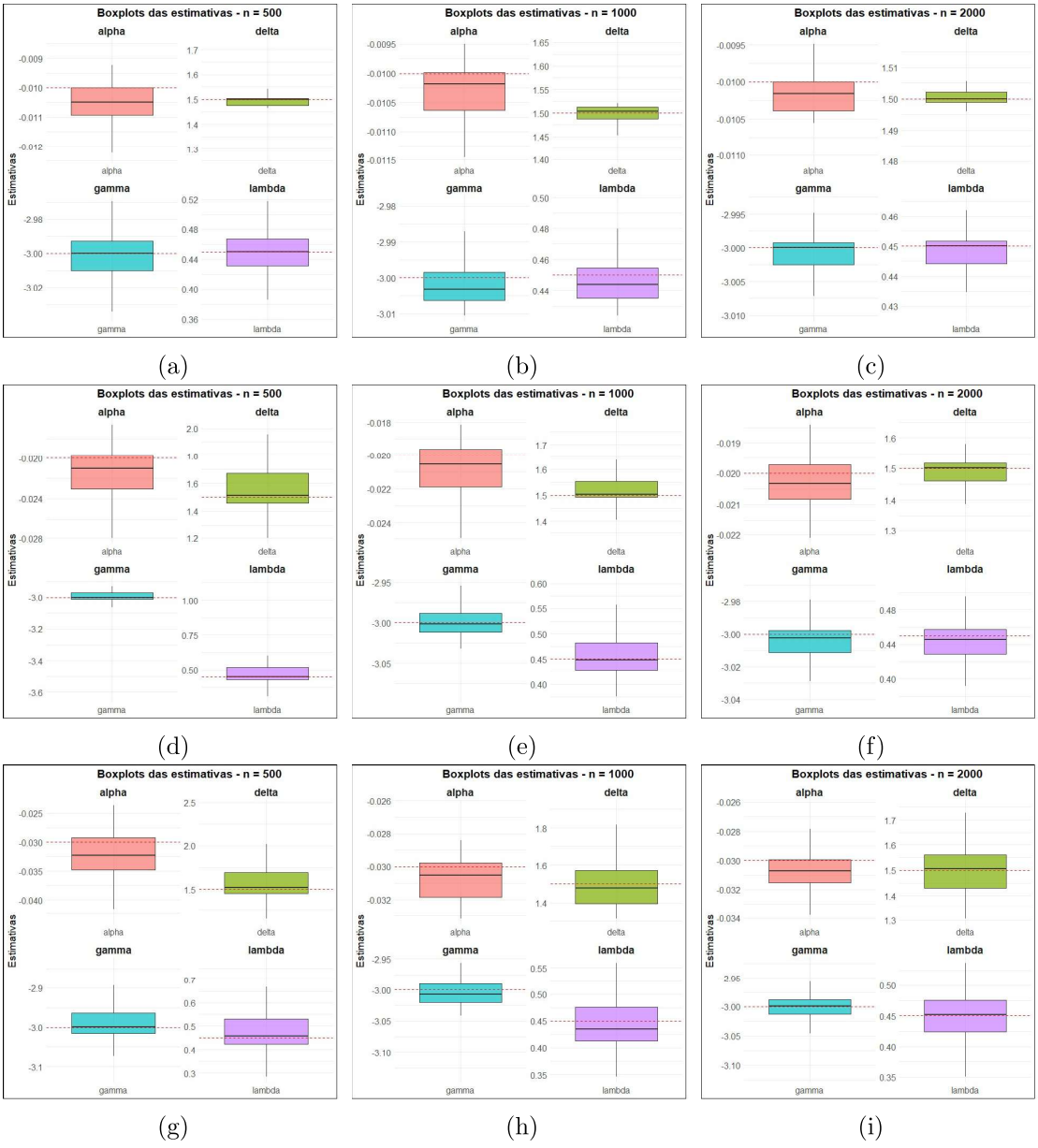


Figura 11: Boxplots das estimativas dos parâmetros do modelo GTDEL no cenário 4($\alpha < 0$), considerando três tamanhos amostrais ($n = 500$, 1000 e 2000) sob diferentes proporções de fração de cura: 16% em (a)–(c), 45% em (d)–(f) e 62% em (g)–(i).

Os boxplots complementam a análise numérica apresentada na Tabela 4.4, permitindo uma avaliação visual da distribuição das estimativas. Observa-se que, para uma mesma proporção de cura, a dispersão das estimativas diminui conforme o tamanho amostral aumenta, com os boxplots tornando- se mais centrados em torno dos valores reais dos parâmetros.

Nas subfiguras (a) a (c), correspondentes ao cenário com fração de cura de 16%, os boxplots mostram estimativas bastante concentradas, mesmo para $n = 500$, com melhorias visíveis à medida que n aumenta. Os parâmetros λ, α, γ e δ são bem estimados, com baixa variabilidade.

Já nas subfiguras (d) a (f), para 45% de cura, e (g) a (i), com 62% de cura, observa-se

um aumento na variabilidade das estimativas, sobretudo para os parâmetros γ e δ. Surgem mais valores atípicos, o que indica que a maior proporção de indivíduos curados, associada a um risco reduzido de falha, resulta em uma quantidade menor de eventos observados, diminuindo, assim, a informação disponível para a estimação dos parâmetros do modelo.

Esses resultados reforçam que, embora o desempenho dos estimadores melhore com o aumento do tamanho amostral, a precisão das estimativas tende a ser afetada quando a proporção de cura é elevada, devido à menor quantidade de eventos observados. Ainda assim, os boxplots mostram que os estimadores permanecem centrados nos valores verdadeiros, o que indica consistência do estimador de máxima verossimilhança no modelo GTDEL.

Capítulo 5

Aplicações para o modelo GTDEL a dados reais

Este capítulo apresenta quatro conjuntos de dados que serão utilizados para avaliar o desempenho do modelo GTDEL. Cada conjunto foi cuidadosamente selecionado para demonstrar o desempenho do modelo em diferentes situações, incluindo dados não censurados, como no caso da fibra de vidro; dados censurados, como nos dados de malária; e dados com presença de fração de cura, exemplificados pelos conjuntos de perda dentária e colangite biliar primária.

As análises foram realizadas no software R (R Core Team, 2024), com comparações sistemáticas entre o modelo GTDEL e seus submodelos (GTDL e TDL). Os resultados incluem análises descritivas, estimativas de parâmetros, critérios de informação (AIC/BIC) e avaliações gráficas de adequação.

5.1 Aplicação 1: Resistência de Fibra de Vidro

O conjunto de dados reais apresentado na Tabela 5.1 refere - se a testes experimentais de resistência de fibras de vidro com $1,5$ cm de comprimento, totalizando 63 observações. Esses dados foram coletados pelo National Physical Laboratory, na Inglaterra, e originalmente estudados por Klakattawi *et al.* (2022). O referido estudo utilizou esse conjunto para ilustrar a aplicabilidade de uma nova família de distribuições, proposta com base na combinação da transformação Marshal- Olkin com a família T-X. Os dados foram essenciais para demonstrar a eficácia e a relevância dessa nova generalização estatística.

Tabela 5.1: Resistência de fibras de vidro.

$0,55$	$0,93$	$1,25$	$1,36$	$1,49$	$1,52$	$1,58$	$1,61$	$1,64$
$1,68$	$1,73$	$1,81$	$2,00$	$0,78$	$1,04$	$1,27$	$1,39$	$1,49$
$1,53$	$1,59$	$1,61$	$1,66$	$1,68$	$1,76$	$1,82$	$2,01$	$0,77$
$1,11$	$1,28$	$1,42$	$1,50$	$1,54$	$1,60$	$1,62$	$1,66$	$1,69$
$1,76$	$1,84$	$2,24$	$0,81$	$1,13$	$1,29$	$1,48$	$1,50$	$1,55$
$1,61$	$1,62$	$1,66$	$1,70$	$1,77$	$1,84$	$0,84$	$1,24$	$1,30$
$1,48$	$1,51$	$1,55$	$1,61$	$1,63$	$1,67$	$1,7$	$1,78$	$1,89$

O conjunto de dados apresentado na Tabela 5.1 tem sido utilizado em estudos estatísticos

relacionados à análise de confiabilidade e resistência de materiais, servindo como referência para testar novos modelos de distribuições. Por exemplo, Smith e Naylor (1987) analisaram o ajuste dos dados de resistência de fibras de vidro utilizando a generalização da distribuição Weibull com três parâmetros. O estudo comparou diferentes modelos de confiabilidade e demonstrou que nova distribuição com três parâmetros proporcionava um ajuste superior em relação às distribuições tradicionais.

De forma semelhante, Aguilar (2017) aplicou o mesmo conjunto de dados para ajustar a generalização da distribuição Gama Exponenciada Poisson Truncada no Zero (GEPTZ), destacando sua flexibilidade e precisão ao modelar os dados de resistência de fibras de vidro.

Mais recentemente, Eghwerido (2022) propôs a distribuição Weibull Frechet Transmutada, explorando suas propriedades teóricas e aplicando-a ao conjunto de dados em questão para demonstrar sua eficácia em modelar falhas de materiais.

A resistência da fibra de vidro, é notável devido às suas propriedades intrínsecas, como alta durabilidade, resistência mecânica e leveza. Essas características tornam esse material extremamente versátil, sendo amplamente empregado em diversas aplicações industriais e de engenharia, incluindo a construção civil, a fabricação de veículos automotivos, além de componentes eletrônicos e aeronáuticos.

Estudos sobre a resistência da fibra de vidro ajudam a desenvolver materiais mais eficientes, aumentando sua longevidade e segurança em diferentes setores industriais.

A Tabela 5.2 apresenta algumas das principais medidas descritivas do conjunto de dados, permitindo a compreensão de características centrais, dispersão e valores extremos. Tais medidas fornecem informações relevantes para resumir e interpretar os dados observados, identificando padrões e avaliando a variabilidade presente, o que é fundamental para análises mais aprofundadas e apoiar a tomada de decisões.

Tabela 5.2: Medidas descritivas das resistências das fibras de vidro.

Medida	**Valor**
Mínimo	0,550
Mediana	1,590
Média	1,507
Máximo	2,240
Variância	0,104
Desvio Padrão	0,323

As medidas descritivas apresentadas na Tabela 5.2 indicam que o valor mínimo de resistência pode estar associado à presença de fibras com defeitos ou variações na qualidade do material. A média é ligeiramente inferior a mediana, o que sugere uma leve assimetria à esquerda na distribuição dos dados.

O valor máximo observado revela a existência de fibras com resistência elevada. De modo geral, a distribuição parece assimétrica, com tendência à concentração de valores acima da média. A diferença entre o mínimo e o máximo é significativa, indicando variabilidade na qualidade ou nas condições das fibras testadas.

A variância e o desvio padrão evidenciam uma dispersão moderada dos valores em torno da média, sinalizando relativa uniformidade entre as amostras. Essa consistência é desejável em contextos industriais e de engenharia, nos quais se busca desempenho estável e previsível.

A maioria das amostras apresenta resistência acima de 1,37, o que confirma as propriedades intrínsecas do material, como alta resistência mecânica e versatilidade, características que justificam seu amplo uso em aplicações industriais e estruturais.

Uma forma eficiente de visualizar a distribuição dos dados é por meio do diagrama de extremos e quartis, conhecido como boxplot (Figura 1). Esse gráfico facilita a identificação de valores atípicos (outliers) e permite observar a assimetria da distribuição por meio da posição

da mediana em relação às bordas da caixa. Além disso, o boxplot é uma ferramenta eficaz para comparar distribuições de diferentes grupos, tornando-o útil em análises comparativas. Como destacam Albert *et al.* (2017), a simplicidade e a clareza do boxplot tornam-no uma escolha ideal para obter uma visão inicial da dispersão e da tendência central dos dados.

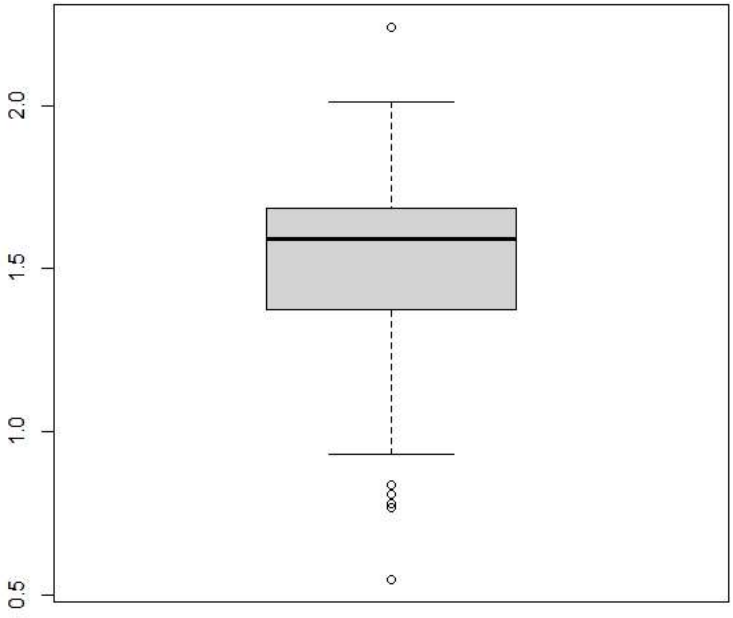


Figura 1: Boxplot dos dados referente a Tabela 5.1.

A Figura 1 exibe o boxplot dos dados, evidenciando que a mediana não está centralizada na caixa, mas deslocada em direção ao terceiro quartil, o que sugere uma assimetria negativa. Além disso, observa-se a presença de outliers em ambos os extremos, com maior concentração no extremo inferior. Esse padrão indica a existência de uma cauda mais longa à esquerda, caracterizando uma distribuição assimétrica com maior variabilidade para valores menores de resistência. Diante dessa distribuição dos dados, apresentados na Tabela 5.1, foram ajustados três modelos estatísticos (GTDEL, GTDL e TDL) com objetivo de avaliar a adequação dos parâmetros estimados e comparar a qualidade do ajuste aos dados. A Tabela 5.3 resume essas análises, apresentando as estimativas dos parâmetros, seus erros-padrão e os critérios de informação AIC e BIC, que auxiliam na comparação entre os modelos e na escolha do mais adequado para descrever os dados.

Tabela 5.3: Estimativas, erros-padrão, AIC e BIC para os parâmetros dos modelos ajustados.

Modelos	**Parâmetro**	**Estimativa**	**Erro-Padrão**	**AIC**	**BIC**
GTDEL	λ	$6,02$	$1,39$	$28,92$	$37,92$
	α	$23,03$	$0,21$		
	γ	$-39,26$	$0,06$		
	δ	$0,15$	$0,02$		
GTDL	λ	15.72	$9,37$	$31,96$	$38,39$
	α	$4,62$	$0,70$		
	γ	$-8,64$	$0,82$		
TDL	α	$6,39$	$1,37$	$98,69$	$102,98$
	γ	$-6,61$	$1,34$		

A Tabela 5.3 mostra que, de acordo com os critérios de informação AIC e BIC, o modelo GTDEL apresenta o melhor ajuste aos dados, quando comparado com seus submodelos GTDL e TDL. Além disso, os erros-padrão associados às estimativas dos parâmetros do modelo GTDEL foram, em sua maioria, menores em relação aos dos demais modelos, o que indica maior precisão nas estimativas obtidas.

A baixa variância observada nas resistências das fibras de vidro (conforme identificado na análise descritiva) contribui para a estabilidade das estimativas no modelo GTDEL. Esse comportamento é refletido nos baixos erros-padrão obtidos para os parâmetros, mesmo quando suas magnitudes são bastante distintas. Por exemplo, o parâmetro α apresentou uma estimativa elevada (23,03) com erro- padrão de apenas 0,21, enquanto δ teve uma estimativa reduzida (0,15), também com erro-padrão baixo (0,02).

A Figura 2 exibe o histograma de frequências dos dados observados com as fdp ajustadas das distribuições GTDEL, GTDL e TDL.

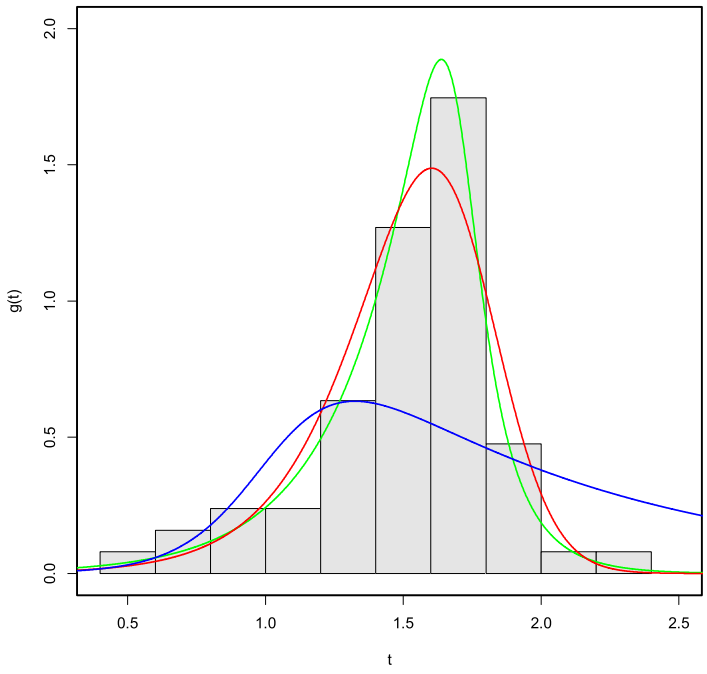


Figura 2: Histograma e funcão densidade de probabilidade estimada para as distribuições GTDEL (verde), GTDL (vermelho) e TDL (azul).

Visualmente, observa-se que a distribuição GTDEL apresenta o melhor ajuste aos dados, pois sua curva acompanha mais de perto a distribuição empírica, especialmente em relação à moda e à assimetria da massa de dados. Esse comportamento indica que o modelo GTDEL consegue capturar com maior precisão as principais características da amostra, oferecendo uma representação mais fiel da variabilidade e estrutura dos dados observados.

Na Figura 3 é exibido o gráfico de probabilidade (*Probability- Probability plot* ou P-P plot) das distribuições ajustadas. Este gráfico compara visualmente as probabilidades empíricas observadas com as probabilidades teóricas esperadas sob o modelo proposto, permitindo avaliar a adequação do ajuste.

O gráfico é definido pela relação:

$$G(t_{(i)}) \quad \text{vs} \quad \frac{i-0{,}375}{n+0{,}25}, \quad i=1,\ldots,n, \tag{5.1}$$

em que $G(\cdot)$ representa a função de distribuição acumulada do modelo teórico, $t_{(i)}$ são os valores amostrais ordenados em ordem crescente e n é o tamanho da amostra. O termo $\frac{i-0{,}375}{n+0{,}25}$ corresponde a uma correção de continuidade proposta por Blom (1958) cuja finalidade é aprimorar a estimativa das probabilidades empíricas. quando os pontos se alinharem próximos à reta de 45° (bissetriz), indica que o modelo teórico descreve adequadamente os dados observados.

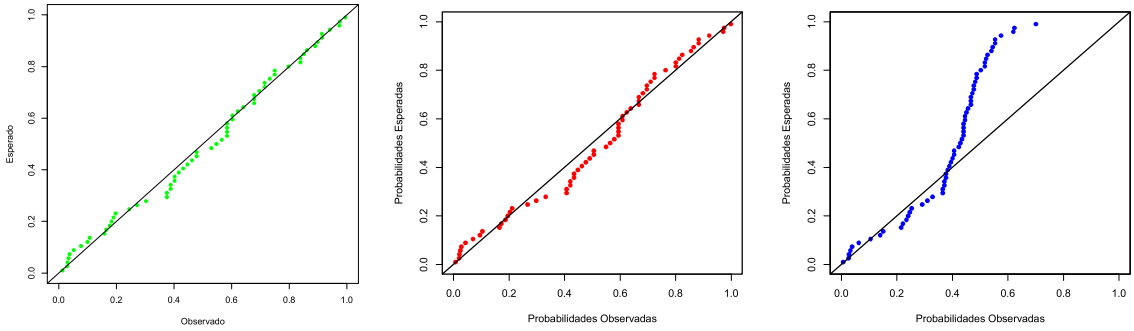


Figura 3: Gráficos de probabilidade das distribuições ajustadas GTDEL (verde), GTDL (vermelho) e TDL (azul).

A análise gráfica da Figura 3 reforça a adequação do modelo GTDEL para descrever os dados apresentados na Tabela 5.1. O gráfico de probabilidade mostra que este modelo apresenta o melhor ajuste, com os pontos mais próximos da linha de referência, o que corrobora os resultados numéricos obtidos por meio dos critérios AIC e BIC. Além disso, essa adequação é consistente com a análise do histograma e da função de densidade, conforme ilustrado na Figura 2.

Entre os modelos alternativos, o modelo GTDL demonstra um ajuste razoável, porém apresenta desvios mais evidentes, especialmente nas extremidades da distribuição. Por outro lado, o modelo TDL exibe um ajuste insatisfatório, com discrepâncias significativas ao longo de toda a distribuição dos dados.

Dessa forma, a escolha do modelo GTDEL como o mais apropriado se sustenta tanto pela análise gráfica quanto pelos critérios estatísticos. Os resultados indicaram um bom ajuste do modelo aos dados, proporcionando estimativas coerentes dos parâmetros e refletindo adequadamente a estrutura dos tempos de falha observados.

5.2 Aplicação 2: Dados de malária

O conjunto de dados apresentados na Tabela 5.4 é proveniente de um estudo experimental descrito em Colosimo e Giolo (2021). O experimento foi conduzido no Centro de Pesquisas René Rachou, Fiocruz (MG), com o objetivo de avaliar a eficácia da imunização pela malária.

A resposta de interesse no estudo foi o tempo decorrido entre a infecção pela malária e a morte do camundongo, medido em dias. Esse tempo reflete a sobrevivência dos camundongos em condições experimentais. O período de acompanhamento foi limitado a 30 dias, sendo as observações dos indivíduos que sobreviveram além desse prazo tratadas como censuradas à direita.

Cada grupo reflete diferentes condições experimentais, permitindo uma análise detalhada das diferenças nos tempos de sobrevivência dos camundongos.

Tabela 5.4: Tempos, em dias, observados no estudo da malária.

Grupos	**Total**	**Tempos de sobrevivência**
Grupo 1	(16)	7, 8, 8, 8, 8, 12, 12, 17, 18, 22, 30+, 30+, 30+, 30+, 30+, 30+
Grupo 2	(15)	8, 8, 9, 10, 10, 14, 15, 15, 18, 19, 21, 22, 22, 23, 25
Grupo 3	(13)	8, 8, 8, 8, 8, 8, 9, 10, 10, 10, 11, 17, 19

Na Tabela 5.5 são apresentadas algumas das principais medidas descritivas do conjunto de dados, como média, mediana, desvio padrão, variância, e os valores mínimos e máximos observados separados por grupo. A análise dessas medidas permite explorar as diferenças entre os tempos de sobrevivência dos camundongos que chegaram ao evento de interesse (morte) e aqueles cujos tempos foram censurados, fornecendo informações importantes sobre a heterogeneidade do conjunto de dados.

Tabela 5.5: Resumo descritivo dos tempos de sobrevivência dos camundongos por grupo, com separação entre dados censurados e não censurados.

Grupo	**Estatística**	**Sem censura**	**Com censura**
GRUPO 1	Quantidade de observações	10	6
	Média (dias)	12,00	–
	Mediana (dias)	10,00	–
	Desvio padrão (dias)	5,27	–
	Valor mínimo (dias)	7	30+
	Valor máximo (dias)	22	30+
GRUPO 2	Quantidade de observações	15	–
	Média (dias)	15,93	–
	Mediana (dias)	15,00	–
	Desvio padrão (dias)	5,95	–
	Valor mínimo (dias)	8	
	Valor máximo (dias)	25	
GRUPO 3	Quantidade de observações	13	–
	Média (dias)	10,31	–
	Mediana (dias)	9	–
	Desvio padrão (dias)	3,59	
	Valor mínimo (dias)	8	–
	Valor máximo (dias)	19	–

Analisando a Tabela 5.5, observa-se que os resultados da análise descritiva mostram diferenças marcantes entre os tempos de sobrevivência observados nos grupos experimentais.

No Grupo 1, composto por 16 camundongos imunizados contra a malária 30 dias antes da infecção e coinfectados com esquistossomose, os dados foram classificados em duas categorias: observações completas (sem censura) e censuradas. Das 16 observações, 10 correspondem a tempos observados até a ocorrência do evento, enquanto 6 são censuradas à direita, com tempo registrado como 30 dias. Para os dados sem censura, a média dos tempos foi de 12 dias, com mediana de 10 dias, desvio padrão de aproximadamente 5,29 dias, valor mínimo de 7 dias e máximo de 22 dias. Já os dados censurados indicam que o evento de interesse não foi observado até 30 dias em 6 indivíduos, sendo 30 o tempo máximo de acompanhamento registrado para esses casos.

Por outro lado, no Grupo 2, formado por 15 camundongos infectados apenas pela malária, todos os indivíduos chegaram ao evento de interesse (morte) antes do término do acompanha-

mento. O tempo médio de sobrevivência foi de 15,93 dias, com uma mediana de 15 dias e menor variabilidade em comparação ao Grupo 1. A ausência de coinfecção com a esquistossomose parece ter contribuído para tempos de sobrevivência relativamente mais homogêneos, ainda que inferiores aos observados no Grupo 1, onde a imunização desempenhou um papel protetor.

No Grupo 3, composto por 13 camundongos infectados tanto pela malária quanto pela esquistossomose, mas sem imunização prévia, também não houve censura, com todos os camundongos alcançando o evento de interesse. Este grupo apresentou o menor tempo médio de sobrevivência, e a menor variabilidade, com desvio padrão de 3,59 dias. A combinação de coinfecção com a ausência de imunização resultou em maior vulnerabilidade, levando a tempos de sobrevivência mais curtos.

Esses resultados destacam o papel da imunização no aumento do tempo médio de sobrevivência e na proteção parcial dos indivíduos no Grupo 1, enquanto a ausência de coinfecção no Grupo 2 proporcionou tempos intermediários de sobrevivência. No entanto, a coinfecção no Grupo 3 teve um impacto severo, resultando nos menores tempos de sobrevivência e indicando a necessidade de estratégias mais eficazes para combater a interação de múltiplas infecções.

A curva de Kaplan-Meier, Figura 4, complementa a análise descritiva ao visualizar dinamicamente as probabilidades de sobrevivência entre os grupos.

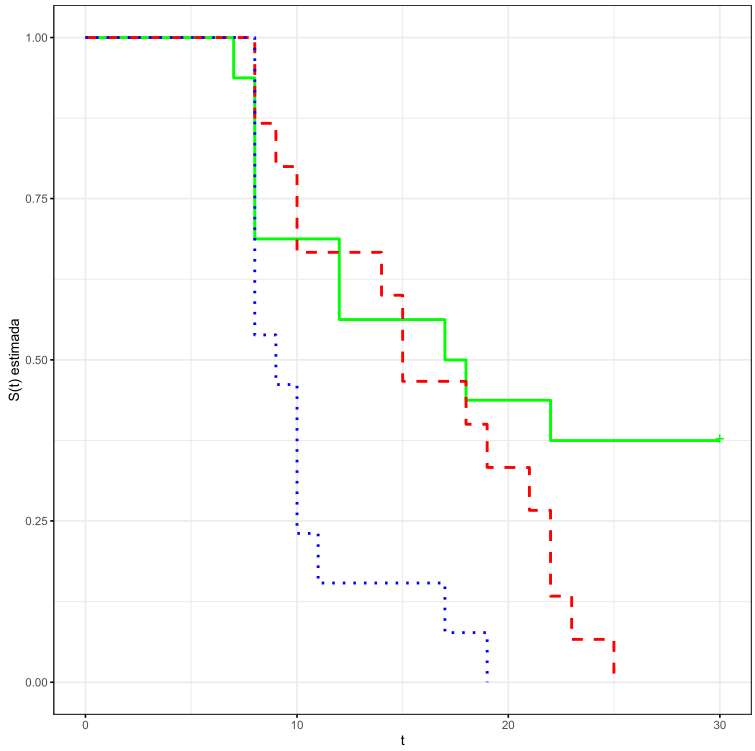


Figura 4: Probabilidade de Sobrevivência estimada por Kaplan-Meier para os dados de malária referentes aos *grupos* 1(*verde*), 2(*vermelho*), e *grupo* 3(*azul*).

As curvas de sobrevivência estimadas por meio do estimador de Kaplan-Meier estão mostradas para os três grupos na Figura 4. Pode-se notar que para o grupo 1, a curva Kaplan-Meier mostra uma função de sobrevivência que decai de forma mais lenta, indicando maior probabilidade de sobrevivência ao longo do tempo. No grupo 2, a curva apresenta uma queda intermediária em relação ao grupo 1, já que todos os indivíduos chegaram ao evento de interesse. Por outro lado, para o grupo 3, a curva apresenta uma queda acentuada e uniforme, terminando em zero antes do final do período de acompanhamento, evidenciando que todos os indivíduos

chegaram ao evento de interesse (morte) em tempos relativamente curtos.

O teste de *logrank* Mantel *et al.* (1966) é um método não paramétrico utilizado para comparar curvas de sobrevivência entre dois ou mais grupos. Sua principal aplicação é testar a hipótese nula de que não há diferença nas funções de sobrevivência entre os grupos ao longo do tempo. A estatística do teste segue uma distribuição qui-quadrado (χ^2), e um p-valor significativo, ($p < 0,05$) indica evidências para rejeitar a hipótese nula, sugerindo diferenças estatisticamente significativas entre os grupos.

A Tabela 5.6, mostra os resultados do teste *logrank* utilizados para as comparações dos grupos, dois a dois, considerados no estudo de malária.

Tabela 5.6: Resultados do Teste de *Logrank* para Comparação de Grupos

Comparação	χ^2	p-valor
Grupo 1 vs Grupo 2	2.5	0.100
Grupo 1 vs Grupo 3	7.9	0.005
Grupo 2 vs Grupo 3	8.0	0.005

Os resultados do teste de *logrank* (Tabela 5.6) revelam diferenças significativas na sobrevivência entre os grupos 1 e 3 e entre os grupos 2 e 3. Entre os grupos 1 e 2, não há diferença significativa. A diferença entre os grupos 1 e 3 atesta a eficácia da imunização contra a malária mesmo na presença de coinfecção, enquanto a disparidade entre os grupos 2 e 3 evidencia o impacto da mortalidade dos camundongos devido à infecção pela esquistossomose.

Dando continuidade à análise, a Tabela 5.7, apresenta as estimativas dos parâmetros, λ, α, γ e δ juntamente com os critérios de informação AIC e BIC, para as distribuições ajustadas GTDEL, GTDL e TDL. Esses modelos foram ajustados ao conjunto de dados descrito na Tabela 5.4. Essa etapa permite avaliar e comparar a qualidade do ajuste de cada modelo, bem como interpretar os parâmetros estimados, que descrevem os padrões observados no experimento.

Tabela 5.7: Estimativas, erros-padrão, AIC e BIC para os parâmetros dos modelos ajustados.

Distribuição	Parâmetro	Estimativa	Erro-Padrão	AIC	BIC
GTDEL	λ	0,03	0,004		
	α	15,29	2,31	120,53	127,66
	γ	−122,38	14,64		
	δ	0,28	0,044		
GTDL	λ	0,10	0,013		
	α	10,11	1,49	130,59	135,94
	γ	−69,39	8,67		
TDL	α	0,07	0,01		
	γ	−3,60	0,51	147,38	150,95

Os resultados apresentados na Tabela 5.7 indicam que o modelo GTDEL apresentou o menor valor de AIC e BIC em relação aos seus submodelos GTDL e TDL, sendo, portanto, o que melhor ajustou os dados do experimento. A estimativa de λ, indica uma taxa inicial de risco baixa, enquanto α sugere um efeito temporal significativo. O parâmetro γ, com valor negativo elevado, reflete uma relação inversa associada ao risco, e o parâmetro δ confere ao modelo maior flexibilidade.

A Figura 5 apresenta o gráfico do estimador de Kaplan-Meier referente ao Grupo 2 de

camundongos, formado por 15 indivíduos que participaram do experimento para avaliar a eficácia da imunização contra a malária. Esse grupo foi definido com base em uma subdivisão específica do estudo, considerando camundongos expostos a um determinado regime de imunização. O gráfico ilustra a função de sobrevivência, que mede a probabilidade de um indivíduo sobreviver além de um determinado tempo t, ou seja, $S(t) = P(T > t)$.

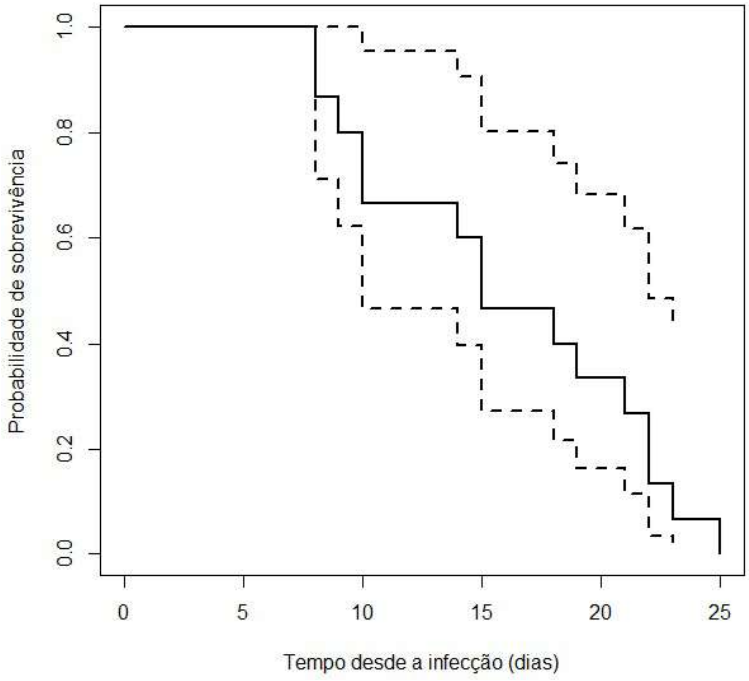


Figura 5: Probabilidade de Sobrevivência estimada por Kaplan-Meier para o grupo 2.

Ao analisar a Figura 5 inicialmente, observa-se que a imunização conferiu proteção eficaz, com uma alta taxa de sobrevivência nos primeiros dias, refletida por uma função de sobrevivência elevada no início do acompanhamento. No entanto, a proteção oferecida diminuiu progressivamente ao longo do tempo, levando a uma queda contínua da função de sobrevivência. Ao final dos 30 dias de observação, todos os camundongos do Grupo 2 sucumbiram ao evento de interesse, evidenciado pela convergência da função de sobrevivência para zero.

Isso indica que não houve sobreviventes ou camundongos resistentes ao longo do estudo, pois todos eventualmente foram suscetíveis à morte causada pela malária. Esses resultados reforçam que, embora a imunização tenha oferecido proteção inicial, ela não foi suficiente para garantir a sobrevivência a longo prazo. Essa limitação sugere a necessidade de explorar estratégias de reforço, como a administração de uma segunda dose ou outras intervenções complementares, com o objetivo de prolongar a proteção conferida pela imunização.

Essas informações são fundamentais para orientar pesquisas futuras sobre a eficácia da vacina, permitindo a investigação de diferentes regimes de imunização e combinações terapêuticas. Além disso, os resultados obtidos no estudo fornecem subsídios importantes para compreender melhor a dinâmica da imunização contra a malária em camundongos e auxiliar no desenvolvimento de estratégias mais eficazes para combater a doença.

5.3 Aplicação 3: Perda Dentária

A perda dentária por doença periodontal aflige a maioria dos adultos ao longo de suas vidas, sendo uma das principais causas de comprometimento funcional e estético da cavidade oral. Diversos fatores influenciam a durabilidade dos tratamentos odontológicos, como idade, presença de doenças sistêmicas (por exemplo, diabetes), hábitos como o tabagismo e variáveis clínicas específicas de cada dente e paciente.

Estudos anteriores utilizaram o conjunto de dados *Teeth* para desenvolver e avaliar modelos estatísticos na previsão da perda dentária. Por exemplo, Hallett *et al.* (2014) aplicaram técnicas de *random forest* para identificar a importância relativa de variáveis na sobrevivência dentária, destacando a relevância da idade, do status de tabagismo e do nível socioeconômico. Mais recentemente, Porndumnernsawat *et al.* (2025) propuseram uma abordagem bayesiana baseada em árvores de sobrevivência com modelos de fragilidade, utilizando extensivamente o banco *Teeth* para analisar padrões de falha dentária em subgrupos da população, como pacientes idosos com e sem diabetes.

O banco de dados *Teeth*, disponível no pacote `MST` do R, contém informações clínicas de 5.336 pacientes com doenças periodontais atendidos na Faculdade de Odontologia da Creighton University School of Dentistry, localizada na cidade de Omaha, no estado de Nebraska, Estados Unidos, entre agosto de 2007 e março de 2013. Os dados incluem o tempo de falha ou censura de 65.228 dentes, sendo 25.331 molares e 39.897 não molares, além de diversas covariáveis clínicas e demográficas. O objetivo principal é modelar o tempo de sobrevivência dos dentes e identificar os fatores associados à falha dentária.

Apesar da disponibilidade de diversas covariáveis clínicas e demográficas no banco de dados *Teeth*, o presente estudo considerou exclusivamente apenas as informações relacionadas ao tempo até a falha ou censura dos dentes, desconsiderando os efeitos das demais variáveis. Essa escolha permite avaliar o desempenho do modelo GTDEL na descrição dos tempos de sobrevivência dentária considerando apenas a estrutura dos dados de tempo e censura, sem influências de variáveis adicionais.

A análise baseada na distribuição GTDEL permite estimar a fração de cura dentária, fornecendo uma medida importante sobre a proporção de dentes que tendem a permanecer viáveis mesmo após longos períodos de acompanhamento. Essa informação é valiosa para a compreensão da dinâmica de falhas em tratamentos odontológicos de longo prazo, especialmente em cenários clínicos complexos, como os relacionados a doenças periodontais. Assim, o modelo proposto contribui com uma ferramenta estatística robusta para apoiar avaliações prognósticas no contexto da saúde bucal.

Na Tabela 5.8 são apresentadas algumas das principais medidas descritivas do tempo de sobrevivência (em anos) do banco de dados *Teeth*, considerando uma subamostra selecionada aleatoriamente. Para esta análise, foi extraída uma amostra composta por 5.336 dentes, mantendo a proporção observada no conjunto completo: 3.735 dentes correspondem a observações censuradas e 1.601 a eventos observados (falhas dentárias). Essa amostragem aleatória visa representar adequadamente a distribuição original dos dados, permitindo avaliar com maior clareza a estrutura de censura e os padrões de falha dentária. A análise descritiva revela padrões distintos entre os grupos: os dentes que não apresentaram falha tendem a permanecer viáveis por períodos mais longos, conforme evidenciado pelas maiores medianas e médias no grupo censurado em comparação ao grupo de falhas. Isso sugere uma possível heterogeneidade na população e reforça a pertinência de modelos que consideram a presença de uma fração de cura. Além disso, observa-se uma maior dispersão dos tempos entre os censurados, refletida pelo maior desvio padrão. Essa variabilidade reflete um padrão comum em dados clínicos ou odontológicos, em que parte dos indivíduos apresenta eventos muito precoces enquanto outros mantêm o evento ausente por longos períodos.

Tabela 5.8: Medidas descritivas dos tempos de sobrevivência para falhas e censuras (em anos)

Medida	**Falhas (event = 1)**	**Censurados (event = 0)**
Mínimo	0,0027	0,0027
Mediana	0,5562	2,1479
Média	1,1381	2,4192
Máximo	5,5808	5,6137
Variância	1,6389	3,3522
Desvio Padrão	1,2802	1,8309

Além disso, observa-se que, no conjunto de dados *Teeth*, cerca de 70% das observações são censuradas, ou seja, referem-se a dentes que permaneceram íntegros até o final do período de acompanhamento. Os 30% restantes correspondem a casos em que houve perda dentária ao longo do estudo.

A Figura 6 apresenta a curva de sobrevivência estimada pelo método de Kaplan-Meier para esses dados.

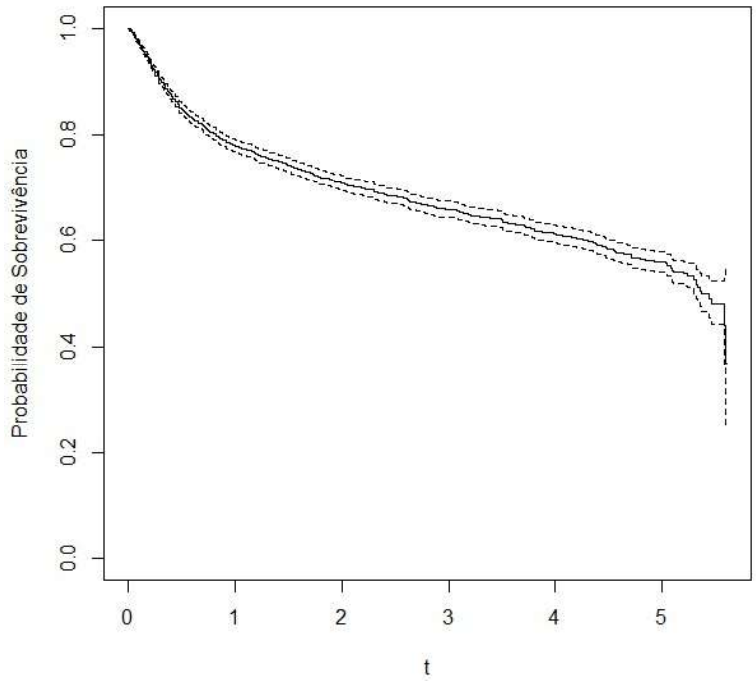


Figura 6: Curva de sobrevivência estimada pelo método de Kaplan-Meier para os dados dentários.

O gráfico de Kaplan-Meier na Figura 6 mostra a probabilidade de um dente permanecer saudável com o passar do tempo, ou seja, sem apresentar eventos como extração, cárie severa ou outras falhas. Esse comportamento é compatível com a hipótese da existência de uma fração de cura e revela aspectos importantes sobre a durabilidade dentária observada no estudo. A taxa de censura observada foi de 70%, o que indica que, para a maioria dos dentes acompanhados, não houve perda dentária durante todo o período do estudo. Esse resultado sugere que a metodologia de acompanhamento pode ter sido relativamente curta para que todas as falhas fossem observadas, que muitos dentes apresentaram boa resistência a problemas que levariam à perda, e que possivelmente houve bons cuidados preventivos na população estudada. Por outro lado, os 30% de eventos registrados (ou seja, perdas dentárias) possibilitam identificar padrões temporais associados às maiores ocorrências de falhas, investigar fatores de risco relacionados a esses casos e reconhecer subpopulações que podem demandar maior atenção odontológica.

A Tabela 5.9 apresenta as estimativas de máxima verossimilhança para os parâmetros

das distribuições GTDEL, GTDL e TDL ajustadas aos dados de sobrevivência dental do banco *Teeth*.

Tabela 5.9: Estimativas, Erro-Padrão, AIC e BIC para os parâmetros dos modelos ajustados.

Distribuição	Parâmetro	Estimativa	Erro-Padrão	AIC	BIC
GTDEL	λ	8,374	9,702	8884,35	8910,76
	α	−0,348	0,034		
	γ	−3,747	1,175		
	δ	0,847	0,031		
GTDL	λ	16,332	16,108	8904,15	8930,48
	α	−0,452	0,024		
	γ	−4,044	1,001		
TDL	α	−0,494	0,026	8921,40	8934,56
	γ	−1,020	0,046		

Os resultados da Tabela 5.9 mostram que a distribuição GTDEL, apresentou os menores valores de AIC e BIC, indicando o melhor ajuste aos dados em comparação com seus submodelos (GTDL e TDL). As estimativas dos parâmetros permitem calcular a fração de cura ρ, que representa a proporção de dentes que não serão perdidos mesmo em tempos prolongados. Para a distribuição GTDEL, essa fração é dada pela Equação (4.4).

A fração de cura estimada foi de 51,12, indicando que aproximadamente 51,12% dos dentes têm probabilidade mínima de falha em cenários de acompanhamento prolongado, isto é, a alta fração sugere que a maioria dos dentes permanece estável.

A Figura 7 apresenta os gráficos de sobrevivência da distribuição GTDEL e do modelo de Kaplan-Meier ajustados aos dados de tempo de sobrevivência dentária.

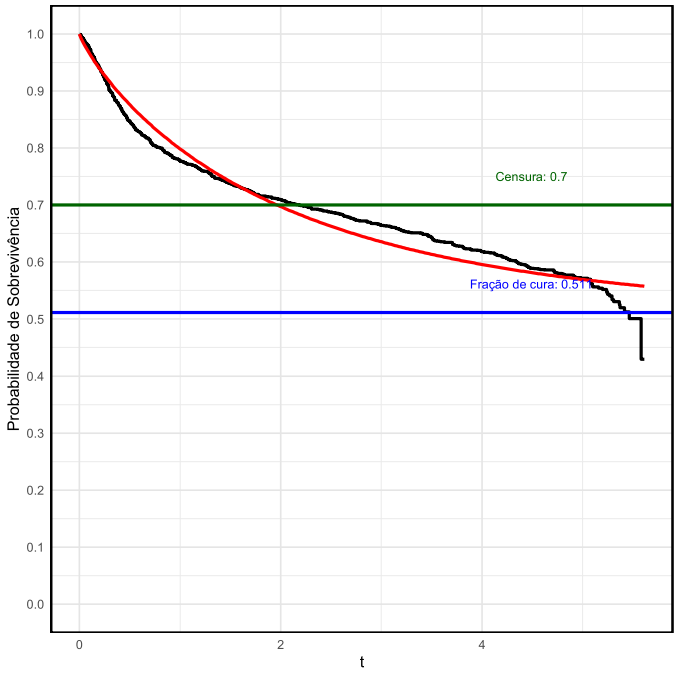


Figura 7: Curvas de sobrevivência dental estimadas pelo método não paramétrico de Kaplan-Meier (preto) vs. distribuição GTDEL (vermelha). A linha azul representa a fração de cura estimada.

A Figura 7 demonstra a eficácia da distribuição GTDEL ao capturar com precisão o comportamento dos dados observado na estimativa de Kaplan-Meier. Percebe-se uma notável concordância entre as curvas, particularmente durante os primeiros 2,5 anos de acompanhamento, o que valida estatisticamente a adequação do modelo proposto. Essa proximidade entre as estimativas paramétrica (GTDEL) e não paramétrica (Kaplan-Meier) reforça a confiabilidade dos resultados obtidos para a análise de sobrevivência dentária.

5.4 Aplicação 4: Colangite Biliar Primária

A colangite biliar primária (CBP) é uma inflamação com fibrose progressiva dos dutos biliares no fígado. Por fim, os dutos são bloqueados, o fígado é tomado por cicatrizes e ocorre o desenvolvimento de cirrose e insuficiência hepática (Lee, 2024). A cirrose hepática é uma causa comum de morte em todo o mundo e foi a 16ª principal causa de morte em 2019.

Neste contexto, a distribuição GTDEL foi ajustada aos dados considerando a presença de uma fração de cura — abordagem que incorpora a possibilidade de que uma parcela dos pacientes não esteja mais sujeita ao risco de óbito. Tal consideração é especialmente relevante em estudos sobre doenças crônicas como a CBP, nas quais alguns indivíduos podem apresentar boa resposta ao tratamento ou formas menos agressivas da enfermidade, sendo, portanto, considerados efetivamente curados (baixo risco de progressão).

Nesta aplicação, foram considerados apenas os dados de tempo até o evento (óbito) e o indicador de censura, sem inclusão de covariáveis explicativas. O objetivo do ajuste do modelo foi obter estimativas dos parâmetros da distribuição GTDEL, de modo a possibilitar o cálculo da fração de cura associada à população estudada.

A cirrose hepática apresenta um curso clínico altamente variável: enquanto alguns pacientes evoluem rapidamente para complicações fatais, outros mantêm a função hepática compensada por anos. A aplicação da distribuição GTDEL permite quantificar a fração de pacientes com comportamento de "cura", capturando com maior precisão a heterogeneidade da evolução clínica desses pacientes.

A Tabela 5.10 apresenta as estimativas dos parâmetros da distribuição GTDEL ajustada aos dados de sobrevivência de pacientes com colangite biliar primária, incluindo erros-padrão e os intervalos de confiança de 95%.

Tabela 5.10: Estimativas, erros-padrão e IC 95% para os parâmetros da distribuição GTDEL defeituosa.

Parâmetro	**Estimativa**	**Erro-Padrão**	**IC 95%**
λ	0,0031	0,0002	[0,0027 ; 0,0035]
α	-0,0009	0,0001	[-0,0011 ; -0,0007]
γ	-0,7863	0,2598	[-1,2952 ; -0,2774]
δ	1,9585	0,1047	[1,7533 ; 2,1637]

Com base nas estimativas obtidas por máxima verossimilhança apresentadas na Tabela 5.10 e mediante substituição na Equação (4.4), estimou- se uma fração de cura de aproximadamente 46,65% para pacientes com diagnóstico de colangite biliar primária. Esse valor teórico, embora menor que a proporção observada de censuras (61,5%), indica que uma proporção significativa dos indivíduos acompanhados não apresentou o evento de interesse (óbito por colangite biliar primária) durante o período de seguimento, evidenciando um prognóstico mais favorável para cerca de metade da população analisada.

É importante destacar que a proporção de censura observada refere-se apenas aos pacientes que não faleceram durante o período de acompanhamento. Essa censura pode ocorrer tanto em indivíduos efetivamente curados quanto naqueles que ainda estavam sob risco e poderiam vir a óbito após o fim do estudo. Além disso, os intervalos de confiança de 95% para os parâmetros

do modelo indicam boa precisão nas estimativas, com todos os limites inferiores e superiores bem definidos.

Capítulo 6

Modelo de Regressão GTDEL

Neste capítulo, apresenta-se o modelo de regressão GTDEL, uma extensão do modelo de regressão GTDL, que incorpora covariáveis ao processo de modelagem da sobrevivência. Inicialmente, são discutidas a formulação do modelo, suas propriedades e a construção da função de verossimilhança sob a presença de covariáveis. Na sequência, descreve-se o procedimento de estimação dos parâmetros do modelo por meio do método da máxima verossimilhança. Por fim, o modelo de regressão GTDEL é aplicado a um conjunto de dados reais provenientes do estudo (CBP), disponível na biblioteca `survival` da linguagem R.

6.1 Modelo de Regressão GTDEL

Seja T uma variável aleatória com distribuição GTDEL, que representa o tempo até a ocorrência do evento de interesse. Conforme estabelecido por Mackenzie (1996), no modelo de regressão GTDL, o parâmetro γ da função densidade de probabilidade (Equação (3.2)), é substituído pela combinação linear $\boldsymbol{X}^\top\boldsymbol{\beta}$, que incorpora o efeito das covariáveis no modelo. Essa parametrização estende-se naturalmente ao modelo GTDEL. O modelo de regressão GTDEL com parâmetros $\lambda > 0$, $\alpha \in \mathbb{R} - \{0\}$ e $\delta > 0$ é caracterizado pelas funções densidade de probabilidade, de distribuição acumulada, de sobrevivência e de risco, respectivamente, como:

$$g(t;\boldsymbol{\theta}) = \lambda\delta\left(\frac{\exp(\alpha t + \boldsymbol{X}^\top\boldsymbol{\beta})}{1+\exp(\alpha t + \boldsymbol{X}^\top\boldsymbol{\beta})}\right) A(t)^{-\frac{\lambda}{\alpha}}\left[1 - A(t)^{-\frac{\lambda}{\alpha}}\right]^{\delta-1}, \tag{6.1}$$

$$G\left(t;\boldsymbol{\theta}\right) = \left[1 - \left(\frac{1+\exp\left(\alpha t + \boldsymbol{X}^\top\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{X}^\top\boldsymbol{\beta}\right)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta},$$

$$S\left(t;\boldsymbol{\theta}\right) = 1 - \left[1 - \left(\frac{1+\exp\left(\alpha t + \boldsymbol{X}^\top\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{X}^\top\boldsymbol{\beta}\right)}\right)^{-\frac{\lambda}{\alpha}}\right]^{\delta}, \tag{6.2}$$

$$h(t;\boldsymbol{\theta}) = \frac{\lambda\delta\left(\dfrac{\exp(\alpha t + \boldsymbol{X}^\top\boldsymbol{\beta})}{1+\exp(\alpha t + \boldsymbol{X}^\top\boldsymbol{\beta})}\right) A(t)^{-\frac{\lambda}{\alpha}}\left[1 - A(t)^{-\frac{\lambda}{\alpha}}\right]^{\delta-1}}{1-\left[1-A(t)^{-\frac{\lambda}{\alpha}}\right]^{\delta}}, \tag{6.3}$$

em que a função auxiliar $A(t)$ é definida por

$$A(t) = \frac{1+\exp(\alpha t + \boldsymbol{X}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta})},$$

em que o vetor de parâmetros $\boldsymbol{\theta} = (\lambda, \alpha, \boldsymbol{\beta}, \delta)^\top$ caracteriza o modelo de regressão GTDEL. O vetor $\boldsymbol{\beta} = (\beta_1,\ \beta_2, \ldots, \beta_p)^\top$ representa os coeficientes (parâmetros) associados às p covariáveis do modelo, e tem dimensão $p \times 1$. Esses coeficientes mensuram a influência de cada covariável sobre o risco do evento de interesse. Já o vetor $\boldsymbol{X}^\top = (x_1,\ x_2, \ldots, x_p)$ contém os valores observados das covariáveis para um indivíduo específico, com dimensão $1 \times p$, podendo incluir, por exemplo, variáveis como idade, tipo de tratamento, entre outras.

A função de risco acumulada do modelo de regressão, conforme MacKenzie e Peng (2014), é dada por:

$$H(t;\boldsymbol{X}) = \int_0^t h(s \mid \boldsymbol{X})ds = \frac{\lambda}{\alpha}\log\left(\frac{1+\exp\{\alpha t + \boldsymbol{X}^\top\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}^\top\boldsymbol{\beta}\}}\right).$$

A expressão $\boldsymbol{X}^\top\boldsymbol{\beta}$, em forma matricial, é dada por:

$$\boldsymbol{X}^\top\boldsymbol{\beta} = \begin{bmatrix} x_1 & x_2 & \ldots & x_p \end{bmatrix}\begin{bmatrix}\beta_1 \\ \beta_2 \\ \vdots \\ \beta_p\end{bmatrix} = \sum_{i=1}^{p} x_i\beta_i.$$

O modelo de regressão GTDEL é denominado modelo de regressão de riscos não proporcionais. De fato, considere dois indivíduos, i e j, sendo $i \neq j$, com covariáveis diferentes $\boldsymbol{X}_i$ e $\boldsymbol{X}_j$, respectivamente.

A razão de riscos entre eles pode ser expressa como:

$$\frac{h(t \mid \boldsymbol{X}_i)}{h(t \mid \boldsymbol{X}_j)} = \frac{\dfrac{\left(\dfrac{\exp(\alpha t + \boldsymbol{X}_i^\top\boldsymbol{\beta})}{1+\exp(\alpha t + \boldsymbol{X}_i^\top\boldsymbol{\beta})}\right) A_i(t)^{-\frac{\lambda}{\alpha}}\left[1 - A_i(t)^{-\frac{\lambda}{\alpha}}\right]^{\delta-1}}{1-\left[1-A_i(t)^{-\frac{\lambda}{\alpha}}\right]^{\delta}}}{\dfrac{\left(\dfrac{\exp(\alpha t + \boldsymbol{X}_j^\top\boldsymbol{\beta})}{1+\exp(\alpha t + \boldsymbol{X}_j^\top\boldsymbol{\beta})}\right) A_j(t)^{-\frac{\lambda}{\alpha}}\left[1 - A_j(t)^{-\frac{\lambda}{\alpha}}\right]^{\delta-1}}{1-\left[1-A_j(t)^{-\frac{\lambda}{\alpha}}\right]^{\delta}}},$$

onde

$$A_k(t) = \frac{1 + \exp(\alpha t + \boldsymbol{X}_k^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}_k^\top \boldsymbol{\beta})}, \quad \text{para } k \in \{i, j\}.$$

Pode-se observar que a razão depende explicitamente do tempo t e das covariáveis $\boldsymbol{X}_i$ e $\boldsymbol{X}_j$.

6.2 Estimação dos parâmetros da regressão GTDEL

Seja $T_i > 0$, $i = 1, 2, \ldots, n$, uma variável aleatória que representa o tempo de falha da i-ésima unidade, e d_i é o indicador de censura, definido como:

$$d_i = \begin{cases} 0, & \text{se o tempo observado for censurado,} \\ 1, & \text{caso contrário.} \end{cases}$$

Assume-se que os tempos de falha T_i são variáveis aleatórias independentes e identicamente distribuídas (IID), com função densidade de probabilidade e sobrevivência dadas, respectivamente, pelas Equações (6.1) e (6.2). Dados $(t_1, d_1), (t_2, d_2), \ldots, (t_n, d_n)$, que consistem em n pares de observações de uma variável aleatória T caracterizada pelo modelo de regressão GTDEL, a função de verossimilhança é definida como:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[g(t_i; \boldsymbol{\theta})\right]^{d_i} \left[S(t_i; \boldsymbol{\theta})\right]^{1-d_i}, \tag{6.4}$$

onde $\boldsymbol{\theta} = (\lambda, \alpha, \boldsymbol{\beta}, \delta)^\top$.

O logaritmo da função de verossimilhança é dado por

$$\begin{aligned} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Bigg\{ d_i \Bigg[& \log(\lambda) + \log(\delta) + \alpha t_i + \boldsymbol{X}_i^\top \boldsymbol{\beta} - \log\left(1 + \exp(\alpha t_i + \boldsymbol{X}_i^\top \boldsymbol{\beta})\right) \\ & - \frac{\lambda}{\alpha} \log A(t_i) + (\delta - 1) \log\left(1 - A(t_i)^{-\lambda/\alpha}\right) \Bigg] \\ & + (1 - d_i) \log\left[1 - \left(1 - A(t_i)^{-\lambda/\alpha}\right)^{\delta}\right] \Bigg\}. \end{aligned}$$

Estimadores de máxima verossimilhança são obtidos derivando-se parcialmente $\ell(\boldsymbol{\theta})$ em relação a $\boldsymbol{\theta}$ e igualando-se o resultado a zero. Esse processo resulta em um sistema de equações não-lineares que deve ser resolvido numericamente.

Para maximização da função de verossimilhança, foi utilizado o método Quasi- Newton, especificamente o algoritmo BFGS (Broyden-Fletcher-Goldfarb- Shanno). Esse método é implementado na função `optim(·, method = "BFGS")` do *software* R (R Core Team, 2024).

6.3 Aplicação

Nesta seção, utilizamos o conjunto de dados provenientes do estudo *pbc* (Primary Biliary Cholangitis), disponível na biblioteca `survival` da linguagem de programação R (R Core Team, 2024). Realiza-se a comparação dos resultados obtidos com os do modelo de regressão GTDL. Essa comparação é feita com base nos parâmetros estimados, nos p- valores associados e no critério AIC, que avalia a qualidade do ajuste dos modelos. Tal análise visa identificar qual dos modelos melhor descreve os dados e captura a dinâmica do risco de óbito ao longo do tempo, considerando as particularidades do conjunto de dados estudado.

Modelos estatísticos de sobrevivência têm sido amplamente utilizados na hepatologia. Contudo, distribuições convencionais frequentemente falham em representar adequadamente a bimodalidade observada empiricamente, com uma fração de pacientes evoluindo rapidamente para óbito e outra apresentando sobrevida significativamente maior. Essa limitação motiva o desenvolvimento de abordagens mais sofisticadas, que incorporem explicitamente a existência de uma fração de longos sobreviventes.

Os registros analisados referem-se a pacientes diagnosticados com colangite biliar primária (CBP), cujas características clínicas e implicações foram detalhadas no Capítulo 5. Trata-se de uma doença hepática crônica que, em estágio avançado, evolui para cirrose e descompensação hepática. Essa condição é reconhecida desde pelo menos 1851 e foi originalmente denominada cirrose biliar primária em 1949. Como a cirrose é característica apenas da fase avançada da doença, uma mudança na nomenclatura para "colangite biliar primária" foi proposta por grupos de defesa dos pacientes em 2014.

A base utilizada na presente análise é proveniente de um ensaio clínico conduzido pela Mayo Clinic sobre *pbc*, realizado entre 1974 e 1984. Um total de 424 pacientes que atenderam aos critérios de elegibilidade foram incluídos no estudo, que avaliou a eficácia do medicamento D-penicilamina, por meio de um ensaio clínico randomizado e controlado por placebo.

O estudo incluiu 424 pacientes, dos quais 312 participaram do ensaio clínico e apresentaram informações mais completas. Outros 112 consentiram em fornecer dados básicos e serem acompanhados quanto à sua sobrevivência, sem participar diretamente do ensaio. Seis casos foram perdidos logo após o diagnóstico, resultando em um total de 418 pacientes com dados disponíveis para análise. No estudo as covariáveis analisadas do conjunto de dados *pbc* foram as seguintes:

- $X_1 = edema$: (0 = ausência, 1 = presença de edema);
- $X_2 = bili$: nível de bilirrubina sérica (mg/dl);
- $X_3 = sexo$: (0 = masculino, 1 = feminino);
- $X_4 = ascites$: (0 = ausência, 1 = presença de ascite);
- $X_5 = hepato$: (0 = ausência, 1 = presença de hepatomegalia);

- $X_6 = spiders$: (0 = ausência, 1 = presença de malformações vasculares);
- $X_7 = albumina$: nível de albumina sérica (g/dl);
- $X_8 = cobre$: Cobre na urina (ug/dia);
- $X_9 = alk.phos$: fosfatase alcalina $U/litro$;
- $X_{10} = trig$: triglicerídeos mg/dl;
- $X_{11} = plaquetas$: contagem de plaquetas;
- $X_{12} = estagio$: estágio histológico da doença (0 = estágio inicial, 1 = estágio avançado);
- $X_{13} = trt$: (0 = placebo, 1 = tratamento);
- $X_{14} = idade$: em anos

O conjunto de dados *pbc* é amplamente utilizado na literatura de análise de sobrevivência para ilustrar e validar técnicas de modelagem, incluindo modelos de regressão, devido à sua relevância clínica. Diferentes abordagens metodológicas têm sido aplicadas a esses dados, demonstrando a evolução das técnicas nesta área.

Por exemplo, Fernandez *et al.* (2020) utilizaram esse banco para investigar a eficácia de diferentes modelos de análise de sobrevivência em pacientes com cirrose biliar primária tratados com D-penicilamina. Os autores concluíram que modelos de aprendizado de máquina, como Random Survival Forest, Gradient Boosting Cox e DeepSurv, apresentaram melhor desempenho do que modelos semi-paramétricos (Cox e Aalen) e paramétricos (Weibull), capturando interações complexas e proporcionando previsões mais precisas.

Rizopoulos (2012) empregou esses dados para investigar a relação entre biomarcadores longitudinais e o tempo até eventos como morte ou recorrência da doença, utilizando modelagem conjunta que integra modelos mistos lineares generalizados para dados longitudinais e modelos de riscos proporcionais de Cox para dados de sobrevivência. Essa abordagem permitiu uma análise mais detalhada da progressão da doença, identificando fatores de risco relevantes.

Já Harrell (2015) utilizou o conjunto *pbc* como estudo de caso para ilustrar técnicas avançadas de modelagem de sobrevivência, incluindo modelos de riscos proporcionais de Cox para identificar fatores de risco que influenciam o tempo até a morte ou transplante de fígado, além de comparar distribuições paramétricas, como Weibull e Exponencial.

Para compreender melhor as características do conjunto de dados, foi realizada uma análise descritiva das covariáveis do conjunto de dados *pbc*. Essa análise permite examinar a distribuição das variáveis contínuas e categóricas, comparando os indivíduos que chegaram ao evento de interesse (óbito) com aqueles cuja observação foi censurada.

As Tabelas 6.1 e 6.2 apresentam um resumo das estatísticas descritivas das variáveis numéricas e categóricas, respectivamente. A Tabela 6.1 exibe medidas como média, mediana, desvio padrão, variância, além dos valores mínimos e máximos observados para cada variável contínua. Essa análise possibilita identificar tendências nos dados, como diferenças na idade, níveis de bilirrubina e contagem de plaquetas entre os indivíduos censurados e não censurados.

Tabela 6.1: Resumo das variáveis numéricas do conjunto de dados *pbc* (Mín = Mínimo; Máx = Máximo; DP = Desvio-padrão; Med = Mediana; Var=Variância).

Variável	**Média**	**Med**	**DP**	**Var**	**Mín - Máx**
$X_2 = bili$	1,77	1,00	2,20	4,85	0,3 - 18,0
X_2^*	5,54	3,20	5,84	34,1	0,3 - 28,0
$X_7 = albumina$	3,58	3,60	0,371	0,138	2,31 - 4,64
X_7^*	3,36	3,40	0,469	0,220	1,96 - 4,52
$X_8 = cobre$	72,5	57,0	64,8	4199,0	4 - 444
X_8^*	135,0	111,0	98,5	9702,0	13 - 588
$X_9 = alk.phos$	1574,0	1132,0	1569,0	2462581,0	289 - 11000
X_9^*	2594,0	1664,0	2677,0	7166910,0	516 - 13900
$X_{10} = trig$	114,0	104,0	51,3	2632,0	33 - 382
X_{10}^*	140,0	122,0	79,3	6282,0	49 - 598
$X_{11} = plaquetas$	266,0	258,0	91,0	8287,0	76 - 539
X_{11}^*	242,0	224,0	108,0	11637,0	62 - 721
$X_{14} = idade$	48,7	48,8	10,4	107,0	26,3 - 78,4
X_{14}^*	53,9	53,5	9,81	96,3	30,9 - 76,7

* : Considerando apenas as observações não-censuradas.

A análise descritiva apresentada na Tabela 6.1, revela diferenças importantes entre os indivíduos censurados e não censurados, permitindo uma melhor compreensão das características associadas à progressão da doença hepática. Por exemplo, observa-se que a idade média dos indivíduos censurados foi de 48,7 anos, enquanto a dos falecidos foi maior, atingindo 53,9 anos, sugerindo uma possível associação entre a idade avançada e um maior risco de óbito.

Um padrão semelhante é identificado nos níveis de bilirrubina sérica, cujo valor médio nos indivíduos censurados foi de 1,77 mg/dL, enquanto nos falecidos foi consideravelmente maior, 5,54 mg/dL, o que indica que níveis elevados dessa substância podem estar relacionados a pior prognóstico. A albumina sérica, por outro lado, apresentou um comportamento oposto: indivíduos censurados apresentaram um nível médio de 3,58 g/dL, enquanto nos falecidos esse valor foi menor, de 3,36 g/dL, o que sugere que uma redução na albumina pode estar associada a um risco maior de óbito.

Outras variáveis também mostram diferenças marcantes entre os dois grupos. O nível médio de cobre na urina foi maior nos falecidos (135 µg/dia) do que nos censurados (72,5 µg/dia), o que pode indicar um agravamento da disfunção hepática nesses pacientes.

A fosfatase alcalina, um marcador de dano hepático, foi significativamente maior nos falecidos, com uma média de 2594 U/L, enquanto os censurados apresentaram um valor médio de 1574 U/L. A contagem de plaquetas também apresentou diferenças relevantes, sendo maior nos censurados (266.000/µL) em comparação com os falecidos (242,000/µL), sugerindo que a queda na contagem plaquetária pode estar relacionada à progressão da doença.

Variáveis como bilirrubina sérica, cobre na urina, fosfatase alcalina e plaquetas apresentam alta variabilidade, com desvios padrão e variâncias elevados. Isso sugere que esses fatores podem variar significativamente entre os indivíduos, possivelmente influenciando a progressão da doença.

A albumina sérica apresenta baixa variabilidade, indicando que os níveis são relativamente consistentes entre os indivíduos, isso significa que a maioria dos valores dessa variável está próxima da média ou da mediana, com pouca dispersão.

Ao observar as estatísticas descritivas, é possível perceber diferenças entre os grupos, o que pode fornecer informações relevantes para a modelagem da sobrevivência desses indivíduos. Essas análises iniciais servem como base para uma avaliação mais aprofundada da relação entre as covariáveis e o tempo até o evento de interesse.

A análise das variáveis categóricas apresentadas na Tabela 6.2 complementa essas observações.

Tabela 6.2: Resumo das variáveis categóricas do conjunto de dados *pbc* (Cens= Censurado; N= número de indivíduos em cada categoria).

Variável	**Categoria**	**Cens (N)**	**Falecidos (N)**	**% Cens**	**% Falecidos**
$X_1 = edema$	Ausente	256	142	20.32	17.36
	Presente	1	19	0.08	2.32
$X_3 = sexo$	Masculino	20	24	4.78	5.74
	Feminino	237	137	56.7	32.8
$X_4 = ascites$	Ausente	186	102	14.76	12.47
	Presente	1	23	0.08	2.81
$X_5 = hepato$	Ausente	115	37	9.13	4.52
	Presente	72	88	5.71	10.76
$X_6 = spiders$	Ausente	149	73	11.83	8.92
	Presente	38	52	3.02	6.36
$X_{12} = stage$	Inicial	88	25	6.99	3.06
	Avançado	167	132	13.25	16.14
$X_{13} = trtx$	Placebo	93	65	7.38	7.95
	Tratamento	94	60	7.46	7.33

A análise das variáveis categóricas apresentadas na Tabela 6.2 permite uma visão detalhada da distribuição dos dados no conjunto *pbc* e da distribuição dos indivíduos censurados e falecidos em cada categoria. Indivíduos que apresentavam edema tiveram uma taxa de mortalidade maior do que aqueles sem essa condição, reforçando a gravidade da retenção de líquidos na progressão da doença.

Quanto ao sexo, observa-se uma maior proporção de indivíduos do sexo feminino no conjunto de dados, isso sugere que a população estudada é majoritariamente feminina, mas a proporção de óbitos entre homens foi ligeiramente maior.

A presença de ascite, um sinal de insuficiência hepática avançada, foi mais frequente nos falecidos, sugerindo que essa condição está associada a um pior prognóstico.

As variáveis hepatoesplenomegalia e presença de *spiders* vasculares, embora tenham apresentado variação entre os grupos, não demonstraram diferenças tão expressivas quanto outras variáveis analisadas.

Além disso, o estágio histológico da doença mostrou diferenças significativas entre os grupos: indivíduos em estágios mais avançados tiveram maior mortalidade em comparação com aqueles nos estágios iniciais, evidenciando que o avanço da fibrose hepática está diretamente relacionado à pior sobrevida.

Por fim, a variável trt (tratamento) mostra uma distribuição equilibrada entre os grupos placebo e tratamento. Isso sugere que o tratamento pode ter um efeito modesto na redução do risco de óbito.

No geral, a análise descritiva destaca que fatores como idade avançada, níveis elevados de bilirrubina e cobre, baixa albumina, presença de ascite e edema, além de um estágio mais avançado da doença, estão associados a um maior risco de óbito.

Dando continuidade à análise, para avaliar a suposição da proporcionalidade dos riscos, utilizou-se o teste baseado nos resíduos de Schoenfeld (1982), implementado pela função cox.zph() no *software* R (R Core Team, 2024). Esse teste avalia a existência de uma relação sistemática entre os resíduos de Schoenfeld e o tempo, sob a hipótese nula de que os coeficientes de risco permanecem constantes ao longo do tempo.

Para cada covariavel, é calculada uma estatística qui-quadrado (χ^2) que avalia essa relação, acompanhada de um p-valor. Se o p- valor for menor que 0,05, rejeita-se a hipótese nula ao nível de significância de 5%, indicando que a covariavel apresenta um efeito não proporcional ao longo do tempo. Esse teste é uma ferramenta fundamental para examinar a validade da suposição de riscos proporcionais no modelo de Cox.

A Tabela 6.3, apresenta os resultados que serviram de base para a construção do modelo final. Durante esse processo, as covariáveis que apresentaram coeficientes não significativos, ou seja, com p-valor superior a 5% ($p > 0,05$), foram removidas. Essa abordagem permitiu simplificar o modelo e aumentar a precisão do ajuste, assegurando uma melhor adequação aos dados. Os resultados obtidos para o teste de Schoenfeld para o conjunto de dados *pbc* encontram-se na Tabela 6.3.

Tabela 6.3: Resultados dos testes de hipóteses associados aos resíduos de Shoenfeld, para checar a pressuposição de riscos proporcionais. (gl = graus de liberdade)

Variável	Estatística χ^2	gl	p-valor
X_1	4,273	1	0,0387
X_2	8,986	1	0,0027
X_3	0,025	1	0,8744
X_4	1,660	1	0,1976
X_5	0,236	1	0,6272
X_6	0,173	1	0,6775
X_7	0,364	1	0,5461
X_8	0,104	1	0,7464
X_9	1,586	1	0,2078
X_{10}	3,552	1	0,0594
X_{11}	2,223	1	0,1360
X_{12}	3,498	1	0,0614
X_{13}	0,178	1	0,6732
X_{14}	2,237	1	0,1347
$GLOBAL$	26,420	15	0,0338

A Tabela 6.3 revela que, ao nível de significância de 5%, as covariávéis edema e bili apresentaram evidências estatisticamente significativas de violação da suposição de riscos proporcionais para essas variáveis.

Além da análise individual das covariáveis, a Tabela 6.3, também apresenta um teste global, que avalia se há evidências de violação da proporcionalidade para o modelo como um todo.

O p-valor do teste global é inferior ao nível de significância de 5%, indicando que, de forma geral, o modelo de Cox pode não ser adequado para o conjunto de dados. Diante desse resultado, a violação da proporcionalidade dos riscos justifica a utilização do modelo de regressão GTDEL como alternativa mais apropriada.

O estudo foi complementado por uma análise gráfica das variáveis que se mostraram significativas no teste analítico. Essa abordagem combinada possibilitou uma avaliação mais detalhada, ao combinar os resultados estatísticos com uma representação visual da evolução dos

riscos ao longo do tempo. O método gráfico adotado consiste em avaliar a proporcionalidade dos riscos com base na estimativa da função de risco acumulado, $H(t)$, ajustada para diferentes grupos de uma covariável.

Na Figura 1, são apresentados os gráficos do logaritmo da função de risco acumulado estimado em relação ao tempo, para as covariáveis *edema* e *bilirrubina*.

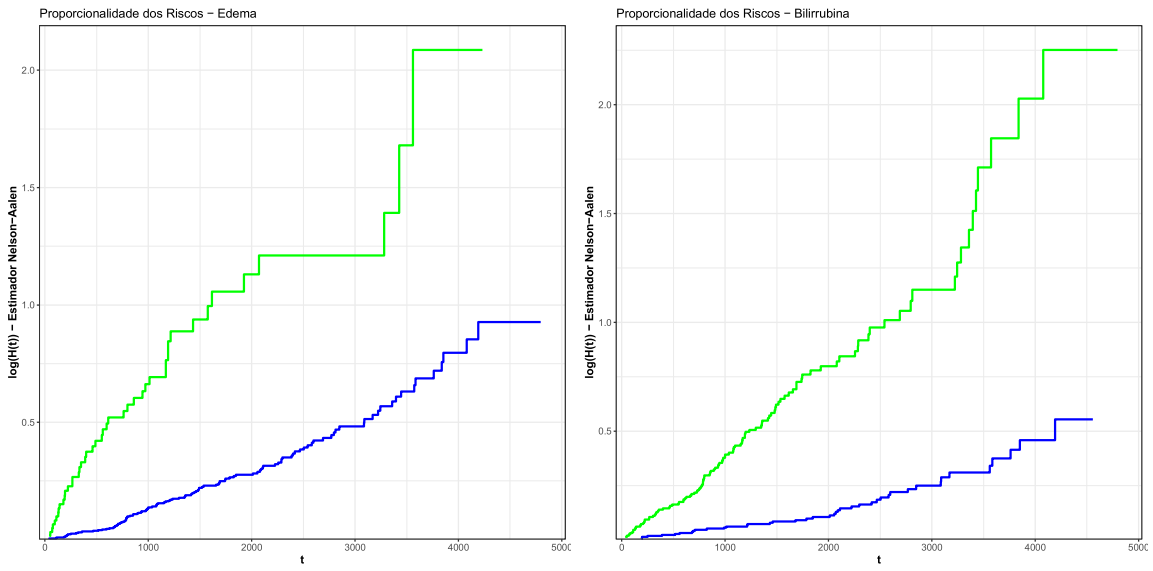


Figura 1: Gráficos do logaritmo da função de risco acumulado estimado em relação ao tempo. À esquerda, a covariável *edema*, com a curva verde representando pacientes sem edema e a curva azul representando pacientes com edema. À direita, a covariável *bili*, com a curva verde indicando pacientes com valores acima da mediana e a curva azul aqueles com valores abaixo da mediana.

Observa-se na Figura 1, que as curvas de log(H(t)) associadas à covariável *edema* apresentam trajetórias divergentes entre os grupos, indicando riscos não proporcionais. Essa dinâmica temporal demonstra que o efeito da presença de edema sobre a mortalidade não é uniforme: seu impacto varia em intensidade conforme o tempo evolui, tornando-se mais crítico em fases específicas.

Da mesma forma, as curvas relativas à covariável *bilirrubina* também indicam ausência de proporcionalidade dos riscos. A divergência observada entre os grupos com níveis acima e abaixo da mediana reforça que níveis elevados de bilirrubina estão fortemente associados a um risco aumentado de óbito, com um efeito que varia ao longo do tempo. Esse comportamento evidencia a necessidade de modelos que considerem a dependência temporal para uma análise mais adequada e precisa dos dados.

Embora as curvas apresentadas na Figura 1 não se cruzem o que representaria uma violação extrema da suposição, elas exibem um comportamento divergente ao longo do tempo, afastando- se de maneira não paralela. Essa falta de paralelismo sugere que a razão de risco não permanece constante, evidenciando graficamente a não proporcionalidade dos riscos.

Para ilustrar a proposta, realizou-se a investigação do tempo até o óbito de pacientes com colangite biliar primária, tendo como variáveis explicativas *edema* e *bilirrubina*. Os resultados, expressos pelas estimativas dos parâmetros obtidas via máxima verossimilhança, estão resumidos na Tabela 6.4 para os modelos de regressão GTDEL e GTDL.

Tabela 6.4: Estimativas, erros-padrão=EP, intervalos de confiança=IC, p-valores e AIC para os parâmetros dos modelos de regressão GTDEL e GTDL.

Modelo	Parâmetro	Estimativa	EP	IC (95%)	p-valor	AIC
GTDEL	λ	0,0006	0,0001	[0,0004; 0,0008]	-	2989,802
	α	-0,0005	0,0001	[-0,0007; -0,0003]	-	
	δ	1,7028	0,1857	[1,3388; 2,0668]	0,0001	
	β_1 (edema)	2,2417	1,0954	[0,0950; 4,3880]	0,0407	
	β_2 (bilirrubina)	0,4081	0,1162	[0,1800; 0,6360]	0,0004	
GTDL	λ	0,0003	-	[0,0003; 0,0004]	0,0001	3010,952
	α	-0,0003	0,0001	[-0,0005; 0,0001]	0,0616	
	β_1 (edema)	3,1208	1,5040	[0,1730; 6,0680]	0,0381	
	β_2 (bilirrubina)	0,4031	0,1402	[0,1260; 0,6800]	0,0040	

Os resultados apresentados na Tabela 6.4 indicam que todos os parâmetros estimados no modelo de regressão GTDEL são estatisticamente significativos ao nível de 5%. O parâmetro α, com valor negativo e estatisticamente significativo, indica que o risco de óbito diminui ao longo do tempo. Tal comportamento pode ser observado em situações clínicas nas quais pacientes que superam os estágios iniciais mais críticos de uma doença passam a apresentar menor probabilidade de morte em momentos posteriores.

As covariáveis selecionadas no modelo, edema e bilirrubina sérica, também apresentaram significância estatística. A estimativa de β_1, associada à variável edema, foi de 2,2417, com intervalo de confiança de 95% entre 0,0950 e 4,3880 e p-valor de 0,0407. Esse resultado indica que a presença de edema está associada a um aumento no risco de óbito. Especificamente, pacientes com edema apresentam um risco de morte aproximadamente $\exp(2{,}2417) \approx 9{,}41$ vezes maior do que aqueles sem edema.

No caso da bilirrubina, a estimativa de β_2 foi de 0,4081, com intervalo de confiança entre 0,1800 e 0,6360 e p-valor de 0,0004, evidenciando forte significância estatística. Isso sugere que valores elevados de bilirrubina sérica estão associados a um maior risco de óbito. Para cada unidade adicional de bilirrubina, estima-se um aumento de aproximadamente $\exp(0,4081) \approx 1,50$, ou seja, um acréscimo de 50% no risco de morte. Esses resultados reforçam a importância clínica dessas duas variáveis e justificam sua inclusão no modelo final.

Por outro lado, ao considerar o modelo de regressão GTDL, que se caracteriza por incorporar riscos não proporcionais com dependência temporal, seria esperado que o parâmetro α, responsável por expressar o efeito do tempo sobre o risco, desempenhasse um papel central na modelagem. No entanto, a estimativa obtida para α não apresentou significância estatística ao nível de 5%, com um p-valor de 0,0616. Esse resultado indica que, para os dados analisados, não há evidências suficientes para confirmar que o risco de óbito varia significativamente ao longo do tempo, o que pode comprometer a adequação do modelo GTDL à realidade observada.

A análise dos intervalos de confiança para os parâmetros do modelo GTDL revela que λ, β_1 (edema) e β_2 (bilirrubina) são estatisticamente significativos, indicando influência direta e relevante sobre o risco de óbito. Em contrapartida, o parâmetro α não se mostra significativo, sugerindo que sua inclusão no modelo pode ser desnecessária ou mesmo inadequada neste contexto. Diante disso, o modelo pode ser simplificado, removendo-se o parâmetro α ou reconsiderando seu papel na estrutura da função de risco.

A comparação dos valores de AIC evidencia que o modelo de regressão GTDEL apresenta desempenho superior em relação ao GTDL. Essa diferença reforça a maior capacidade do GTDEL em capturar as nuances da influência temporal e dos efeitos das covariáveis sobre o risco de óbito. Assim, conclui-se que o modelo de regressão GTDEL é a abordagem mais adequada para

representar os dados do estudo, especialmente em contextos nos quais a variação do risco ao longo do tempo desempenha um papel relevante.

Adicionalmente, a comparação entre o modelo de regressão GTDEL e o modelo de regressão GTDL foi realizada por meio do Teste da Razão de Verossimilhança (TRV). Esse teste tem como objetivo verificar se a inclusão do parâmetro adicional δ resulta em um ajuste significativamente melhor aos dados. A hipótese nula considerada foi

$$H_0 : \delta = 1$$

que implica que o parâmetro extra não é necessário, contra a hipótese alternativa

$$H_1 : \delta \neq 1.$$

A estatística do teste é dada por:

$$\Lambda = 2(\ell_{\text{GTDEL}} - \ell_{\text{GTDL}}) = 2(-1489{,}901 + 1501{,}476) = 23{,}15,$$

em que ℓ_{GTDEL} e ℓ_{GTDL} representam, respectivamente, os valores das log-verossimilhanças obtidos com os modelos completo e reduzido. Esse valor é comparado ao quantil da distribuição qui-quadrado, cujo número de graus de liberdade corresponde à diferença entre a quantidade de parâmetros dos modelos considerados. Como o modelo de regressão GTDEL possui cinco parâmetros e o modelo GTDL possui quatro, resulta em um grau de liberdade.

O valor crítico da distribuição qui-quadrado com um grau de liberdade ao nível de significância de 5% é:

$$\chi^2_{1,,0,95} \approx 3{,}84.$$

Como $\Lambda = 23{,}15 > 3{,}84$, rejeita-se a hipótese nula. Portanto, conclui-se que o modelo de regressão GTDEL fornece um ajuste significativamente melhor aos dados analisados.

Essa conclusão também pode ser visualmente corroborada pela Figura 2, na qual se observa que a curva ajustada pelo modelo GTDEL acompanha de forma mais próxima a curva empírica de Kaplan-Meier, especialmente nos períodos intermediários e finais de acompanhamento. Já o modelo GTDL tende a superestimar a probabilidade de sobrevivência ao longo do tempo, destacando sua limitação em representar adequadamente a heterogeneidade dos riscos ao longo do tempo.

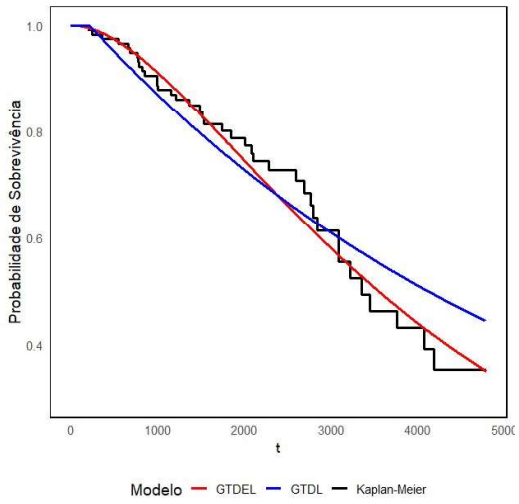


Figura 2: Curvas de sobrevivência estimadas pelos modelos GTDL e GTDEL, comparadas à curva de Kaplan-Meier.

Capítulo 7

Conclusões

Neste trabalho, introduziu-se uma nova distribuição com quatro parâmetros, denominada Modelo Logístico Generalizado Estendido Dependente do Tempo (GTDEL), que inclui como submodelos a família Logística Tempo-Dependente e o Modelo Logístico Generalizado Dependente do Tempo. Essa nova distribuição amplia as possibilidades de modelagem, oferecendo maior flexibilidade para capturar a complexidade dos dados de sobrevivência, incluindo cenários com fração de cura.

Uma das principais contribuições teóricas foi a caracterização da distribuição GTDEL defeituosa, que ocorre para valores do parâmetro $\alpha < 0$. Nessa configuração, a função de sobrevivência não converge para zero, mas sim para um valor assintótico $\rho \in (0, 1)$, representando a proporção de indivíduos imunes ao evento de interesse. Essa propriedade permite modelar adequadamente situações em que parte da população nunca experimenta o evento, o que é particularmente relevante em estudos clínicos e industriais. A expressão matemática da fração de cura foi derivada e discutida, destacando-se a influência do parâmetro δ na magnitude de ρ e evidenciando sua importância no controle da proporção de curados. Também se discutiu o comportamento da fração de cura ρ sob diferentes configurações paramétricas, fornecendo uma base sólida para sua interpretação prática.

Para fins de inferência, adotou-se o método da máxima verossimilhança para a estimação dos parâmetros do modelo GTDEL. Essa abordagem se mostrou apropriada não apenas por suas propriedades assintóticas desejáveis — como consistência e eficiência —, mas também por sua aplicabilidade em diferentes cenários de modelagem.

O desempenho dos estimadores foi avaliado por meio de um estudo de simulação baseado no Método de Monte Carlo, considerando métricas como média, viés e raiz do erro quadrático médio (REQM). Os resultados indicaram que, à medida que o tamanho amostral aumenta, as estimativas convergem para os valores verdadeiros dos parâmetros, com redução progressiva do viés e do REQM, o que reforça a consistência do método de estimação adotado.

A aplicação da distribuição proposta a dados reais demonstrou um ajuste de qualidade superior em comparação com seus submodelos, destacando seu potencial para a análise de dados de sobrevivência. Esse resultado reforça a relevância da nova distribuição como uma ferramenta poderosa para modelar fenômenos complexos, especialmente em situações onde os modelos tradicionais falham em capturar a dinâmica dos dados.

Como sugestão para trabalhos futuros, destaca-se a investigação de novas estratégias de estimação, bem como o desenvolvimento de ferramentas diagnósticas, como a análise de resíduos adaptada ao modelo GTDEL. Tais ferramentas poderão avaliar a adequação do modelo aos dados empíricos, contribuindo para seu aprimoramento e para uma interpretação mais robusta dos resultados.

Outra linha promissora de investigação diz respeito à relação entre censura e fração de cura em dados de sobrevivência. Observações empíricas indicam que a curva de Kaplan-Meier pode apresentar um aparente platô, sugerindo uma estabilização da sobrevivência; no entanto,

esse comportamento nem sempre reflete a presença real de indivíduos curados. Isso ocorre, por exemplo, quando há registros de eventos após o último tempo censurado, fazendo com que a curva continue a decrescer até o final do acompanhamento. Nesses casos, torna-se evidente que a proporção de censurados não deve ser confundida com a fração de cura.

Enquanto a censura representa indivíduos que não apresentaram o evento até o fim do estudo, ela inclui casos que poderiam eventualmente experimentar o desfecho, caso o seguimento fosse prolongado. A curva de Kaplan-Meier reflete essa incerteza, ao passo que a proporção de censura é uma medida descritiva, sem caráter inferencial. Nesse sentido, estudos futuros poderiam explorar métodos capazes de distinguir com maior precisão entre censura informativa, censura não informativa e cura verdadeira, considerando o impacto do tempo de censura na identificação da fração de cura em modelos paramétricos e semiparamétricos.

Por fim, este trabalho contribui para o avanço da modelagem de sobrevivência ao propor e avaliar um modelo que captura a dinâmica do risco ao longo do tempo e incorpora a possibilidade de fração de cura. Os resultados obtidos demonstram a relevância do modelo GTDEL e abrem caminho para novas pesquisas que possam aprimorar sua aplicação e interpretação. Assim, espera-se que este estudo sirva como base para futuras investigações na área de análise de sobrevivência e modelagem estatística.

Referências Bibliográficas

Afify, A. e Abdellatif, A. (2020). The extended burr xii distribution: properties and applications. *Journal of Nonlinear Sciences and Applications*, **13**(3), 133–146.

Aguilar, G. A. S. (2017). *Família distribuição gama exponenciada*. Dissertação de mestrado, Universidade Estadual Paulista (Unesp), Presidente Prudente, SP.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723.

Al-Tawarah, Y. e MacKenzie, G. (2003). A non-ph accelerated hazard model for analyzing clinical trial data.

Albert, J., Glickman, M. E., Swartz, T. B., e Koning, R. H. (2017). *Handbook of Statistical Methods and Analyses in Sports*. CRC Press, Boca Raton, FL.

Ali, S., Ali, S., Shah, I., Siddiqui, G. F., Saba, T., e Rehman, A. (2020). Reliability analysis for electronic devices using generalized exponential distribution. *IEEE Access*, **8**, 108629–108644.

Aranda-Ordaz e J, F. (1983). An extension of the proportional-hazards model for grouped data. *Biometrics*, **39**(1), 109–117.

Balka, J., Desmond, A. F., e McNicholas, P. D. (2009). Review and implementation of cure models based on first hitting times for wiener processes. *Lifetime data analysis*, **15**, 147–176.

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

Blagojevic-Bucknall, M. e MacKenzie, G. (2004). Ph and non-ph frailty models for multivariate survival data. *Statistical Modelling*, **4**(1), 69–91.

Blom, G. (1958). *Statistical Estimates and Transformed Beta-Variables*. Ph.D. thesis, Almqvist & Wiksell.

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

Bolfarine, H. e Sandoval, M. C. (2001). *Introdução à Inferência Estatística*. Sociedade Brasileira de Matemática, Rio de Janeiro, 2 edition.

Boyer, C. B. e Merzbach, U. C. (2011). *A history of mathematics*. John Wiley & Sons, New York, NY.

Cancho, V. G., Rodrigues, J., e de Castro, M. (2011). A flexible model for survival data with a cure rate: a bayesian approach. *Journal of Applied Statistics*, **38**(1), 57–70.

Carrasco, J. M., Ortega, E. M., e Cordeiro, G. M. (2008). A generalized modified weibull distribution for lifetime modeling. *Computational Statistics & Data Analysis*, **53**(2), 450–462.

Casella, G. e Berger, R. L. (2024). *Statistical Inference*. CRC Press, Boca Raton, 3 edition.

Castro, R. V. O., Araújo Júnior, C. A., Leite, H. G., Castro, A. F. N. M., Nogueira, G. S., e Costa, L. S. (2016). Função gama generalizada para descrever a distribuição diamétrica de um povoamento de eucalipto. *Floresta*, **46**(3), 355–364.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency ofin chronic disease incidence. *Biometrika*, **65(1)**, 141–151.

Colosimo, E. (1997). A note on the stratified proportional hazards model. *International Journal of Mathematical and Statistical Sciences*, **6**, 201–210.

Colosimo, E. A. e Giolo, S. R. (2021). *Análise de Sobrevivência Aplicada*. Editora Blucher, São Paulo, 2 edition.

Cordeiro, G. M., Ortega, E. M., e Silva, G. O. (2011). The exponentiated generalized gamma distribution with application to lifetime data. *Journal of statistical computation and simulation*, **81**(7), 827–842.

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

de Oliveira, A. G. e de Fátima Leite, M. (2024). Doenças do fígado são causa de 3% das mortes e custam r$ 300 milhões anuais ao país. *UFMG Notícias*. Acesso em: 24 maio 2025.

Dey, T., Mandal, A., e Chakraborty, S. (2020). An overview of study design and statistical considerations: A practical overview and reporting strategies for statistical analysis of survival studies. *Chest Journal*, **158**(1S), 539–548.

Eghwerido, J. T. (2022). The transmuted weibull frechet distribution: Properties and applications. *Reliability: Theory & Applications*, **17**(4 (71)), 453–468.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, **38**(4), 1041–1046.

Fernandez, C., Chen, C. S., Gaillard, P., e Silva, A. (2020). Comparação experimental de modelos semi-paramétricos, paramétricos e de aprendizado de máquina para análise de tempo até evento por meio do índice de concordância. *arXiv preprint arXiv:2003.08820*.

Gieser, P. W., Chang, M. N., Rao, P., Shuster, J. J., e Pullen, J. (1998). Modelling cure rates using the gompertz model with covariate information. *Statistics in medicine*, **17**(8), 831–839.

Gomes, M. I. (2024). The portsea and a few personal scientific achievements. *Communications in Mathematics*, **32**(3), 329–391.

Gompertz, B. (1825). Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. *Philosophical transactions of the Royal Society of London*, pages 513–583.

Gove, J. H. (2017). A demographic study of the exponential distribution applied to uneven-aged forests. *Forestry: An International Journal of Forest Research*, **90**(1), 18–31.

Gupta, R. D. e Kundu, D. (2001). Exponentiated exponential family: an alternative to gamma and weibull distributions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **43**(1), 117–130.

Hallett, K. B., Radford, K. J., Seow, W. K., e Martins, S. (2014). Random forests for tooth loss prediction: importance of socioeconomic status. *Statistical Modelling*, **14**(5), 439–455.

Harrell, F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics. Springer, Cham, 2 edition.

Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika*, **71**(1), 75–83.

Ibrahim, J. G., Chen, M.-H., e Sinha, D. (2013). *Bayesian Survival Analysis*. Springer Science & Business Media, New York.

Ihaka, R. e Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**(3), 299–314.

James, B. R. (2004). *Probabilidade: um curso em nível intermediário*. Instituto de Matemática Pura e Aplicada, CNPq.

Kalbfleisch, J. D. e Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2 edition.

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

Klakattawi, H., Alsulami, D., Elaal, M. A., Dey, S., e Baharith, L. (2022). A new generalized family of distributions based on combining marshal-olkin transformation with tx family. *PloS one*, **17**(2), e0263673.

Lee, M.-L. T. e Whitmore, G. A. (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science*, **21**(4), 501–513.

Lee, T. H. (2024). Colangite biliar primária (cbp). Revisado e atualizado em fevereiro de 2024.

Longo, A. J., Sampaio, S. C., Queiroz, M. M., e Suszek, M. (2006). Uso das distribuições gama e log-normal na estimativa de precipitação provável quinzenal. *Varia Scientia*, **6**(11), 107–118.

Louzada, F., Cuminato, J. A., Rodriguez, O. M. H., Tomazella, V. L., Milani, E. A., Ferreira, P. H., Ramos, P. L., Bochio, G., Perissini, I. C., Junior, O. A. G., *et al.* (2020). Incorporation of frailties into a non-proportional hazard regression model and its diagnostics for reliability modeling of downhole safety valves. *IEEE Access*, **8**, 219757–219774.

Louzada-Neto, F., Cremasco, C. P., e MacKenzie, G. (2010). Tsampling-based inference for the generalized time-dependent logistic hazard model. *Journal of Statistical Theory and Applications*, **9**, 169–184.

Louzada-Neto, F., Mackenzie, G., Cremasco, C. P., e Ferreira-Silva, P. H. (2011). On the interval estimation of the parameters of a generalized time-dependent logistic model. *Revista Brasileira de Biometria*, **29**, 512–519.

Mackenzie, G. (1996). Regression models for survival data: The generalized time-dependent logistic family. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **45**(1), 21–34.

MacKenzie, G. e Peng, D. (2014). *Statistical Modelling in Biostatistics and Bioinformatics: selected papers*. Springer.

MacKenzie, G., Blagojevic-Bucknall, M., Al-tawarah, Y., e Peng, D. (2003). A comparison of non-ph gamma frailty models. *In: Proceedings of the 17th International Workshop in Statistical Modelling*, **18**, 39–44.

Magalhães, M. N. (2006). *Probabilidade e Variáveis Aleatórias*. Edusp, São Paulo, 1 edition.

Maller, R. A. e Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley Series in Probability and Statistics. Wiley, New York.

Mantel, N. *et al.* (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, **50**(3), 163–170.

McKeague, I. W. e Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, **81**(3), 501–514.

Milani, E. A. (2011). *Modelo logístico generalizado dependente do tempo com fragilidade*. Dissertação de mestrado, Universidade Federal de São Carlos (UFSCar), São Carlos, SP.

Mudholkar, G. S., Srivastava, D. K., e Kollia, G. D. (1996). A generalization of the weibull distribution with application to the analysis of survival data. *Journal of the American Statistical Association*, **91**(436), 1575–1583.

Nocedal, J. e Wright, S. (2006). *Numerical Optimization*. Springer Science & Business Media, 2 edition.

Oliveira, L., Santos, L., Fabio, L., Ferreira, P., e Carrasco, J. (2023). Análise de resíduos para o modelo logístico generalizado dependente do tempo. *Trends in Computational and Applied Mathematics*, **24**, 635–658.

Pascoa, M. A. R. d. (2012). *Extensões da distribuição gama generalizada: propriedades e aplicações*. Ph.D. thesis, Universidade de São Paulo.

Peng, Y., Qi, X., e Guo, X. (2016). Child–pugh versus meld score for the assessment of prognosis in liver cirrhosis: a systematic review and meta-analysis of observational studies. *Medicine*, **95**(8), e2877.

Porndumnernsawat, P., Frank, T., e Ingsrisawang, L. (2025). On a bayesian multivariate survival tree approach based on three frailty models. *Scientific Reports*, **15**(1), 12017.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramires, T. G., Ortega, E. M., Lemonte, A. J., Hens, N., e Cordeiro, G. M. (2020). A flexible bimodal model with long-term survivors and different regression structures. *Communications in Statistics-Simulation and Computation*, **49**(10), 2639–2660.

Ramos, P. L. (2014). *Aspectos computacionais para inferência na distribuição gama generalizada*. Dissertação de mestrado, Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP), Presidente Prudente, SP.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman & Hall/CRC Biostatistics Series. CRC Press, Boca Raton.

Robert, C. P. e Casella, G. (2010). *Introducing Monte Carlo Methods with R*, volume 18. Springer.

Rocha, R., Nadarajah, S., Tomazella, V., e Louzada, F. (2016). Two new defective distributions based on the marshall–olkin extension. *Lifetime Data Analysis*, **22**(2), 216–240.

Rodrigues, J., de Castro, M., Cancho, V. G., e Balakrishnan, N. (2009). Com–poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, **139**(10), 3605–3611.

Santos, M. R. d., Achcar, J. A., e Martinez, E. Z. (2017). Bayesian and maximum likelihood inference for the defective gompertz cure rate model with covariates: an application to the cervical carcinoma study. *Ciência e Natura*, **39**(2), 244–258.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**(1), 239–241.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, **6**, 461–464.

Scudilio, J., Calsavara, V. F., Rocha, R., Louzada, F., Tomazella, V., e Rodrigues, A. S. (2019). Defective models induced by gamma frailty term for survival data with cured fraction. *Journal of Applied Statistics*, **46**(3), 484–507.

Shakil, M. e Kibria, B. G. (2009). Exact distributions of the linear combination of gamma and rayleigh random variables. *Austrian Journal of Statistics*, **38**(1), 33–44.

Shawky, A. e Bakoban, R. (2008). Bayesian and non-bayesian estimations on the exponentiated gamma distribution. *Applied Mathematical Sciences*, **2**(51), 2521–2530.

Smith, R. L. e Naylor, J. C. (1987). A comparison of maximum likelihood and bayesian estimators for the three-parameter weibull distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **36**(3), 358–369.

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

Struthers, C. A. e Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, **73**(2), 363–369.

Therneau, T. M. e Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer, New York.

Tibshirani, R. J. C. (1983). A family of proportional and additive hazard models for survival data. *Biometrics*, **39**(1), 141–147.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, **18**, 293–297.

Whttmore, G. (1979). An inverse gaussian model for labour turnover. *Journal of the Royal Statistical Society: Series A (General)*, **142**(4), 468–478.