**UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"**

Instituto de Biociências - Câmpus de Botucatu

Curso de Física Médica

# MODELOS DE REGRESSÃO EM ANÁLISE DE SOBREVIVÊNCIA: TEORIA E SIMULAÇÃO DE DADOS COM APLICAÇÃO NA ÁREA DA SAÚDE

Laura Sauer Maschetti

Orientador: Profª. Drª. Miriam Harumi Tsunemi

Botucatu – SP

2025

**UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"**

Instituto de Biociências - Câmpus de Botucatu

Curso de Física Médica

# MODELOS DE REGRESSÃO EM ANÁLISE DE SOBREVIVÊNCIA: TEORIA E SIMULAÇÃO DE DADOS COM APLICAÇÃO NA ÁREA DA SAÚDE

Laura Sauer Maschetti

Trabalho de Conclusão de Curso apresentado à Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Instituto de Biociências - Câmpus de Botucatu, como parte dos requisitos para obtenção do título de Bacharel em Física Médica.

Orientador: Profª. Drª. Miriam Harumi Tsunemi

Botucatu – SP

2025

M395m

Maschetti, Laura Sauer
Modelos de regressão em análise de sobrevivência : Teoria e simulação de dados com aplicação na área da saúde. / Laura Sauer Maschetti. -- , 2025
49 p.

Trabalho de conclusão de curso ( - ) - Universidade Estadual Paulista (UNESP), Instituto de Biociências, Botucatu,
Orientadora: Miriam Harumi Tsunemi

1. Análise de Sobrevivência. 2. Simulação computacional. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

LAURA SAUER MASCHETTI

# MODELOS DE REGRESSÃO EM ANÁLISE DE SOBREVIVÊNCIA

Teoria e Simulação de dados com aplicação na área da saúde

Trabalho de Conclusão de Curso, apresentado a Universidade Estadual Paulista, como parte das exigências para a obtenção do título de Bacharel, do curso de Graduação em Física Médica.

Botucatu, 18 de Novembro de 2025.
(data da defesa)

BANCA EXAMINADORA

Miriam H. Tsunemi

Prof ª. Miriam Harumi Tsunemi
Departamento de Biodiversidade e Bioestatística, UNESP Botucatu

Prof. Paulo Fernando de Arruda Mancera
Departamento de Biodiversidade e Bioestatística, UNESP Botucatu

Rogerio Antonio de Oliveira

Prof. Rogerio Antonio de Oliveira
Departamento de Biodiversidade e Bioestatística, UNESP Botucatu

# Agradecimentos

Gostaria de expressar meus sinceros agradecimentos a todos que contribuíram direta ou indiretamente para a realização deste projeto de pesquisa. Primeiramente, agradeço a minha orientadora, Miriam Harumi Tsunemi, pelo apoio constante, paciência e valiosas orientações ao longo de todo o desenvolvimento do projeto. Seu conhecimento, dedicação e incentivo foram fundamentais para a execução e conclusão deste trabalho.

Agradeço também ao Departamento de Biodiversidade e Bioestatística do Instituto de Biociências do Câmpus de Botucatu, por disponibilizar a infraestrutura e as condições adequadas à realização do projeto, possibilitando que todas as etapas fossem desenvolvidas de forma eficiente e organizada.

Gostaria de agradecer especialmente à Coordenação do Programa Institucional de Bolsas de Iniciação Científica (PIBIC) da Reitoria, pelo fomento financeiro e institucional que tornou possível a realização deste projeto. O apoio recebido pelo programa foi essencial para o desenvolvimento das atividades de pesquisa, permitindo a aplicação prática dos conhecimentos teóricos adquiridos durante a graduação.

Por fim, deixo meu agradecimento mais afetivo e sincero à minha família, em especial aos meus pais, Luciana e Sandro, e à minha irmã, Mariana, pelo amor e incentivo incondicionais. Ao meu namorado, João Victor, que foi a calmaria em meio à tempestade da graduação, obrigado por ser meu porto seguro e minha fonte inesgotável de apoio e carinho.

# Resumo

Este trabalho analisou e comparou modelos de regressão aplicados à análise de sobrevivência utilizando dados simulados e aplicações em saúde. Foram revisados conceitos fundamentais, como tempo de falha, censura e funções probabilísticas, além do uso do estimador de Kaplan-Meier. A etapa experimental consistiu em simulações no RStudio com as distribuições exponencial, Weibull e log-normal, ajustadas aos modelos paramétricos correspondentes. A avaliação dos modelos foi feita por máxima verossimilhança, log-verossimilhança e resíduos de Cox-Snell. Os resultados indicaram melhor flexibilidade dos modelos Weibull e log-normal, enquanto o modelo exponencial mostrou adequação apenas quando a taxa de risco é constante. O estudo reforça a utilidade dos modelos paramétricos de regressão na análise de sobrevivência em contextos de saúde.

**Palavras-chave:** análise de sobrevivência. regressão paramétrica. Kaplan-Meier. Weibull. log-normal. modelo exponencial. resíduos de Cox-Snell. máxima verossimilhança.

# Abstract

This study analyzed and compared regression models applied to survival analysis using simulated data and applications in the health field. Fundamental concepts such as failure time, censoring, and associated probability functions were reviewed, along with the use of the Kaplan-Meier estimator. The experimental stage involved simulations in RStudio using exponential, Weibull, and log-normal distributions, later fitted to their corresponding parametric models. Model performance was evaluated through maximum likelihood, log-likelihood, and Cox-Snell residuals. Results showed that the Weibull and log-normal models offered greater flexibility, while the exponential model was suitable only when the hazard rate remains constant. The study reinforces the relevance of parametric regression models as robust tools for analyzing survival times in health-related contexts.

**Keywords:** survival analysis. parametric regression. Kaplan-Meier. Weibull. log-normal. exponential model. Cox-Snell residuals. maximum likelihood.

# Lista de Figuras


4.1 Gráficos das sobrevivências estimadas por Kaplan-Meier *versus* as estimada pelos modelos exponencial, Weibull e log-normal, respectivamente. . . . . 26

4.2 Gráfico de sobrevivência *versus* o tempo em meses por Kaplan-Meier e os modelos paramétricos. . . . . 27

4.3 Análise dos resíduos de Cox-Snell do modelo de regressão exponencial ajustado. 28

4.4 Análise dos resíduos de Cox-Snell do modelo de regressão Weibull ajustado. . . 29

4.5 Análise dos resíduos de Cox-Snell do modelo de regressão log-normal. . . . . 29

4.6 Estimador de Kaplan-Meier com diferenciação da remissão em homens e mulheres e valor de p = 0.03, dado pela análise com teste "logrank". . . . . 31

4.7 Estimador de Kaplan-Meier com diferenciação da remissão em homens e mulheres e valor de p = 0.03, dado pela análise com teste "logrank". . . . . 32

# Lista de Tabelas

4.1 Comparação entre os valores verdadeiros e estimados dos parâmetros dos modelos apresentados. . . . . 30
4.2 Coeficientes estimados ($\beta_0$, ($\beta_1$ e $\beta_2$) para o modelo exponencial e valor de p. . 33
4.3 Coeficientes estimados ($\beta_0$, ($\beta_1$ e $\beta_2$) para o modelo Weibull e valor de p. . . . 33
4.4 Coeficientes estimados ($\beta_0$, ($\beta_1$ e $\beta_2$) para o modelo log-normal e valor de p . . 34

# Lista de Algoritmos

A.1 Simulação dos dados - Primeira Análise . . . . . 38
A.2 Ajustes das simulações com Máxima Verossimilhança . . . . . 38
A.3 Comparação dos parâmetros simulados e os estimados por máxima verossimilhança. . . . . 39
A.4 Gráficos ponto a ponto e Log Verossimilhança . . . . . 41
A.5 Gráficos para avaliação dos modelos . . . . . 42
A.6 Gráfico dos Resíduos e seu algoritmo. . . . . 42
A.7 Simulação dos dados - Segunda Análise . . . . . 44
A.8 Plot gráfico Kaplan-Meier e Teste Logrank. . . . . 45
A.9 Modelos de Regressão Paramétrico. . . . . 46
A.10 Gráfico dos Modelos Regressão Paramétrico. . . . . 46
A.11 Coeficientes dos modelos aplicados aos dados simualdas. . . . . 48

# Sumário


**1 Introdução** . . . . . **1**

**2 Revisão Bibliográfica** . . . . . **3**

2.1 Conceitos Básico na Análise de Sobrevivência . . . . . 3

2.2 Representação dos Dados de Sobrevivência . . . . . 4

2.3 Especificando o Tempo de Sobrevivência . . . . . 5

2.3.1 Função de Sobrevivência . . . . . 5

2.3.2 Função de Taxa de Falha ou Risco . . . . . 5

2.3.3 Função de Taxa de Falha Acumulada . . . . . 6

2.3.4 Tempo Médio e Vida Média Residual . . . . . 6

2.3.5 Relação entre as Funções . . . . . 7

2.4 Estimador de Kaplan-Meier e Teste *Logrank* . . . . . 7

2.5 Modelos Paramétricos em Análise de Sobrevivência e Estimação dos Parâmetros 10

2.5.1 Distribuição Exponencial . . . . . 10

2.5.2 Distribuição de Weibull . . . . . 11

2.5.3 Distribuição Log-normal . . . . . 12

2.5.4 Método de Máxima Verossimilhança . . . . . 12

2.6 Modelos de Regressão para Dados de Sobrevivência . . . . . 17

2.6.1 Modelo Linear para Dados de Sobrevivência . . . . . 17

2.7 Adequação do Modelo Ajustado . . . . . 22

2.8 Critério de Informação de Akaike (AIC) . . . . . 23

**3 Material e Métodos** . . . . . **24**

3.1 Simulação no RStudio . . . . . 24

3.1.1 Simulação dos Modelos Probabilísticos . . . . . 24

3.1.2 Simulação do conjunto de Dados com Variáveis Independentes . . . . . 25

**4 Resultados e Discussão** . . . . . **26**
4.1 Simulação dos modelos probabilísticos . . . . . 26
4.2 Simulação do conjunto de Dados com Variáveis Independentes . . . . . 30
4.2.1 Coeficientes dos modelos paramétricos . . . . . 32

**5 Conclusão** . . . . . **35**

**A Anexos** . . . . . **38**
A.1 Algoritmos Utilizado na construção dos Resultados da Primeira Análise . . . . 38
A.2 Algoritmos Utilizado na construção dos Resultados da Segunda Análise . . . . 44

# 1. Introdução

A Análise de Sobrevivência é uma área da Estatística que envolve o estudo de pesquisas as quais há o seguimento de unidades experimentais durante um determinado período. Nestes estudos, há usualmente a perda de informação e neste sentido, a modelagem por meio de Análise de Sobrevivência permite a utilização de informação parcial considerando os casos perdidos como informação censurada. Exemplos de censuras podem ser pacientes que mudaram de cidade ou faleceram por motivos além da doença de interesse.

As principais áreas de aplicação são a médica, biológicas e agrárias. Um exemplo importante de aplicação prática é o estudo de Couchoud *et al.* (2017), que estimou o tempo médio de sobrevivência restrito (RMST) em 15 anos para 67.258 pacientes iniciando terapia de substituição renal (pessoas que começaram hemodiálise, diálise peritoneal ou receberam transplante após entrar em insuficiência renal terminal) na França. O estudo revelou reduções expressivas na expectativa de vida conforme a idade e a presença de comorbidades, de 14,3 anos para pacientes sem diabetes com 18 anos para 1,8 anos aos 90 anos, evidenciando o impacto clínico e a relevância dos métodos de sobrevivência em pesquisas médicas.

Os aspectos mais importantes para as medidas utilizadas são as funções de sobrevivência e de risco, pois indicam o comportamento da sobrevida dos indivíduos em determinada situação. A função de sobrevivência indica a probabilidade do evento de interesse ocorrer além de um tempo específico e a função de risco indica a taxa instantânea de ocorrência do evento dado que o evento não ocorreu até determinado tempo (Colosimo; Giolo, 2006).

Os modelos de regressão permitem avaliar a influência de uma ou mais covariáveis sobre o tempo de sobrevida. Eles são muito úteis para avaliar características de indivíduos ou até comparar diferentes tratamentos. Assim, uma pesquisa pode auxiliar no direcionamento do tratamento de indivíduos com determinada doença, contribuindo na redução da mortalidade prematura, promovendo a saúde e o bem-estar.

A utilização de modelos de Análise de Sobrevivência em muitas áreas do conhecimento cresceu bastante devido ao avanço computacional. Inicialmente, o desenvolvimento de métodos numéricos aliados à evolução da tecnologia propiciou a disseminação das formas de estimação em conjunto ao aumento da capacidade de organização e armazenamento de informações ex-

perimentais de forma digital. Com o desenvolvimento tecnológico, a estimação pontual para intervalos e testes estatísticos de hipótese de funções de sobrevivências tornaram-se viáveis. No estudo desenvolvido por Bewick, Cheek e Ball (2004) temos uma revisão dos métodos de Kaplan-Meier e o teste de *logrank* que foi facilitada pelos avanços tecnológicos.

Neste sentido, destaca-se o programa estatístico R, pois é disponibilizado gratuitamente. Na década de 90, em 1993, os professores Ross Ihaka e Robert Gentleman, do departamento de Estatística da Universidade de Auckland, na Nova Zelândia, desenvolveram esse software. (Ihaka; Gentleman, 1993)

O nome "R"foi escolhido a partir das iniciais dos nomes dos seus criadores. O "R"é um ambiente de software gratuito utilizado para realizar análises estatísticas e criar gráficos. Sua principal característica é sua capacidade de ser compilado e executado em diversos sistemas operacionais, como Windows, Unix e macOS. O RStudio é uma das interfaces para o "R"que busca beneficiar os usuários.

Pesquisadores de várias partes do mundo contribuem constantemente, permitindo o avanço contínuo de técnicas estatísticas e a sua disponibilidade para a sociedade científica. O domínio desta linguagem de programação permite a modelagem e análise de diversos tipos de experimentos, como no estudo de Rai e outros (2022) que se utilizou da linguagem R para auxiliar na obtenção dos resultados por meio de gráficos plotados.

O objetivo, portanto, do presente trabalho é o estudar modelos de regressão em dados de sobrevivência simulados e suas aplicações na área da saúde. Os modelos de regressão probabilísticos utilizados foram o exponencial, Weibull e log-normal. Este estudo foi dividido em duas etapas:

- Estudo sobre os conceitos básicos da Análise de Sobrevivência referentes à técnicas não paramétricas de estimação de curva de sobrevivência, modelos probabilísticos como exponencial, Weibull e log-normal, além dos modelos de regressão paramétricos.
- Utilização do programa estatístico R por meio do RStudio para a execução das análises por simulação, junto com o pacote `survival`.

# 2. Revisão Bibliográfica

Neste capítulo terá a apresentação dos principais conceitos na área de Análise de Sobrevivência Estatística referentes à técnicas não paramétricas de estimação de curva de sobrevivência, modelos probabilísticos como exponencial, Weibull e log-normal, além dos modelos de regressão paramétricos.

## 2.1 Conceitos Básico na Análise de Sobrevivência

A Análise de sobrevivência envolve problemas em que os pacientes são acompanhados e registra-se o tempo até a ocorrência de um evento de interesse, sendo geralmente a variável resposta. Os eventos de interesse são, geralmente, os indesejados e são chamados de "falhas". É crucial defini-los claramente antes do estudo, como no caso de produtos alimentícios, onde o evento seria o momento em que o produto se torna inadequado para consumo. O evento (falha) pode ocorrer por uma ou mais causas. Em ocasiões nas quais os eventos competem entre si, levam o nome na literatura de "riscos competitivos"(Prentice *et al.*, 1978).

- **Tempo de falha:** É o tempo do início da pesquisa até o tempo estimado da ocorrência do evento de interesse. Por exemplo em estudos da área da saúde, pode ser a análise do diagnóstico de alguma doença até a remissão ou até o paciente ficar livre de sintomas da doença ou até mesmo a morte do paciente.

- **Censura:** São situações em que o acompanhamento foi interrompido antes do término do evento analisado. Por exemplo os pacientes estudados mudaram de cidade, o paciente morreu, por exemplo.

Portanto em estudos de análise de sobrevivência há três situações que precisam ser definidas. O tempo inicial, a escala das medidas e o evento de interesse, junto com a censura, devem estar claros para o estudo.

O tempo de início do estudo precisa ser detalhadamente definido. A data de início de tratamento de alguma doença, por exemplo, ou o diagnóstico, podem ser escolhas para o tempo inicial. A escala de medida pode ser considerada o tempo real em dias, meses, etc.

As censuras são observações incompletas, pois é comum que nem todos os indivíduos experimentem o evento de interesse antes do término do estudo. As censuras fornecem informações parciais sobre o tempo até o evento e não devem ser descartadas, pois mesmo incompletas podem trazer informações como o tempo de vida de um paciente e a análise pode resultar em conclusões enviesadas, por conta da omissão da censura nos cálculos estatísticos.

Os principais tipos de mecanismos de censura são:

- **Censura do tipo I**: O estudo termina após um tempo fixo pré-estabelecido.
- **Censura do tipo II**: O estudo termina depois de ter ocorrido o evento de interesse em um número pré-determinado de indivíduos.
- **Censura Aleatória**: É o tipo de censura que mais ocorre na área médica em que alguns indivíduos deixam o estudo antes da falha, seja por desistência, perda de acompanhamento ou morte por outras causas.

Além disso, a censura podem ser: à direita, à esquerda e intervalar. A **censura à direita** é aquela em que o evento de interesse não é observado até o final do estudo e é o caso mais comum em estudo de sobrevivência. A **censura à esquerda** é aquela em que o evento ocorreu antes do início do estudo e a **intervalar** é aquela em que só sabe-se que o evento de interesse ocorreu dentro de algum intervalo de tempo, porém não sabe exatamente quando.

## 2.2 Representação dos Dados de Sobrevivência

Os dados de sobrevivência de um indivíduo $i$ ($i = 1, \ldots, n$) em análise, geralmente, são representados pelo par $(t_i, \delta_i)$, em que $t_i$ é o tempo até a falha ou censura, e $\delta_i$ é a variável que indica se ocorreu uma falha ou uma censura, podendo então ser definido como:

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é um tempo de falha,} \\ 0, & \text{se } t_i \text{ é um tempo censurado.} \end{cases} \tag{2.1}$$

Quando há covariáveis associadas ao $i$-ésimo indivíduo, como, por exemplo, $\mathbf{x}_i = (\text{sexo}_i, \text{idade}_i, \text{tratamento}_i)$, os dados passam a ser representados por $(t_i, \delta_i, \mathbf{x}_i)$ para $i = 1, \ldots, n$. Para os dados de sobrevivência intervalar, inclui-se, a representação $(\ell_i, u_i, \delta_i, \mathbf{x}_i)$ em que $\ell_i$ e

$u_i$ são, respectivamente, os limites inferior e superior do intervalo observado para o $i$-ésimo indivíduo para $i = 1, \ldots, n$.

## 2.3 Especificando o Tempo de Sobrevivência

A variável aleatória $T$ deve ser positiva e contínua, ou seja, a sua função densidade de probabilidade $f(t)$ assume valores entre 0 e 1 para valores positivos de $t$. Há diversas formas de representar a distribuição de probabilidade, além da função de probabilidade $f(t)$, tal como a função de sobrevivência, a função de risco e a função taxa de falha. Nesta seção, serão apresentadas estas medidas que são importantes na área de Análise de Sobrevivência.

### 2.3.1 Função de Sobrevivência

Segundo Colosimo e Giolo (2006), a função de sobrevivência da variável aleatória $T$ é definida como a probabilidade de uma observação não falhar até um certo tempo $t > 0$:

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

Assim, tem-se que:

- $S(0) = 1$, pois no início do estudo, a probabilidade de não ter ocorrido o evento é total.
- $S(t) \to 0$, quando $t \to \infty$, assumindo que eventualmente o evento ocorrerá.

Dessa forma, a função de sobrevivência está relacionada à função de distribuição acumulada $F(t)$ do tempo até o evento que é definida como a probabilidade do indivíduo não sobreviver até o tempo $t$.

### 2.3.2 Função de Taxa de Falha ou Risco

A função de taxa de falha $\lambda(t)$ é uma ferramenta para caracterizar a distribuição do tempo. Ela descreve a forma em que a taxa instantânea de falha varia ao longo do tempo.

A função de taxa de falha de $T$ no tempo $t$, para $t > 0$, é dada por:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.2}$$

Em análise de sobrevivência ela representa o risco instantâneo de falha no tempo $t$, $t > 0$ e é sempre positivo.

### 2.3.3 Função de Taxa de Falha Acumulada

A taxa de falha acumulada do indivíduo no tempo $t$, para $t > 0$, é definida como:

$$\Lambda(t) = \int_0^t \lambda(u)\, du.$$

Esta função tem uma interpretação indireta, mas pode ser vantajoso na avaliação da função de taxa de falha.

### 2.3.4 Tempo Médio e Vida Média Residual

Duas medidas de interesse em análise de sobrevivência são: o tempo médio de vida e a vida média residual. O tempo médio de vida é sobre a expectativa de duração antes que ocorra o evento de interesse. Tal integral representa a área sob a curva $S(t)$, a função de sobrevivência dada por:

$$t_m = \int_0^\infty S(t)\, dt.$$

A vida média residual corresponde a um certo tempo de vida $t$, ou seja, para indivíduos com idade $t$, ela indicará uma duração média do tempo a partir de $t$ e é determinada com a área sob a curva de sobrevivência à direita do tempo $t$ dividida por $S(t)$, dada por:

$$\text{vmr}(t) = \frac{\int_t^\infty (u - t) f(u)\, du}{S(t)} = \frac{\int_t^\infty S(u)\, du}{S(t)},$$

Com $f(\cdot)$ sendo a função de densidade de $T$. Observa-se que $\text{vmr}(0) = t_m$.

### 2.3.5 Relação entre as Funções

Para $T$ uma variável aleatória contínua e não-negativa, tem-se algumas relações matemáticas entre as funções definidas:

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

$$\Lambda(t) = \int_0^t \lambda(u)\, du = -\log S(t),$$

e

$$S(t) = \exp\{-\Lambda(t)\} = \exp\left\{-\int_0^t \lambda(u)\, du\right\}.$$

Essas relações mostram que ao se conhecer uma das funções, por exemplo $S(t)$, obtêm-se as outras como: $F(t)$, $f(t)$, $\lambda(t)$ e $\Lambda(t)$. Outras relações entre essas funções são:

$$S(t) = \frac{\text{vmr}(0)}{\text{vmr}(t)} \exp\left\{\int_0^t \frac{du}{\text{vmr}(u)}\right\}, \quad \lambda(t) = \left(\frac{d\,\text{vmr}(t)}{dt} + 1\right)/\text{vmr}(t).$$

## 2.4 Estimador de Kaplan-Meier e Teste *Logrank*

O estimador de Kaplan-Meier (Kaplan; Meier, 1958) é um dos mais utilizados para a forma não-paramétrica, sendo ele uma adaptação da função de sobrevivência empírica que, com a falta da censura, é definida como:

$$\hat{S}(t) = \frac{\text{número de observações que não falharam até o tempo t}}{\text{número total de observações no estudo}}. \tag{2.3}$$

Na construção do estimador de Kaplan-Meier divide o tempo em intervalos correspondentes aos tempos de falha da amostra. Os limites de cada intervalo terminam no momento em que ocorre uma falha. O processo de estimativa segue uma sequência condicional, ou seja, a probabilidade de sobreviver até um certo tempo depende dos tempos de sobrevivência anteriores.

Considerando as observações $(t_1, \delta_1), \ldots, (t_n, \delta_n)$ de $n$ indivíduos tal que $\delta_i$ é definido como em (2.1). Para construir o estimador de Kaplan-Meier devemos:

- Ordenar os $k$ tempos distintos de falha como $t_1 < t_2 < \cdots < t_k$,

- Utilizar as seguintes notações:
  - $d_j$: número de falhas no tempo $t_j$;
  - $n_j$: número de observações sob risco que não falharam e que não foi censurado até o tempo anterior a $t_j$.

Assim, o estimador de Kaplan-Meier para $S(t)$, é dado por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left( \frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left( 1 - \frac{d_j}{n_j} \right). \tag{2.4}$$

Kaplan e Meier, no artigo original, mostra que a função (2.4) é o estimador de máxima verossimilhança de $S(t)$. A verossimilhança mede o quanto um modelo estatístico explica os dados que observamos. Ela é parecida com probabilidade, porém na probabilidade busca-se saber a chance de algo acontecer com os parâmetros conhecidos, já na verossimilhança, busca-se avaliar a informação de uma amostra para diferentes valores dos parâmetros a partir de um conjunto de dados. A função de verossimilhança para $S(t)$ pode ser escrita então como: (Colosimo; Giolo, 2006).

$$L(S(\cdot)) = \prod_{j=0}^{k} \left\{ \left[ S(t_j) - S(t_{j+}) \right]^{d_j} \prod_{\ell=1}^{m_j} S(t_{j\ell}+) \right\}.$$

Assim, a expressão (2.4) é aquela que maximiza a verossimilhança $L(S(\cdot))$ em que $t_j+$ é o tempo imediatamente após $t_j$ representando uma queda na função $S(t)$ e $t_{j\ell}+$ o tempo de censura, sem precisar assumir um modelo paramétrico fixo.

De modo geral, as principais características do estimador de Kaplan-Meier podem ser resumidas da seguinte maneira (Colosimo; Giolo, 2006):

i) Para amostras grandes, ele não apresenta vícios;

ii) Possui consistência fraca;

iii) Apresenta convergência assintótica para um processo gaussiano;

iv) Corresponde ao estimador de máxima verossimilhança da função $S(t)$.

O teste *Logrank* é muito utilizado para comparar curvas de sobrevivência de dois ou mais grupos, sendo adequado quando a razão das funções de risco é aproximadamente constante (Colosimo; Giolo, 2006, p. 55).

Considerando que a hipótese nula seja verdadeira $H_0$, ou seja, $S_1(t) = S_2(t)$, em que $S_1(t)$ e $S_2(t)$ são as funções de sobrevivência dos grupos 1 e 2, e $t_1 < t_2 < ... < t_k$ são os tempos de falha distintos formados a partir da combinação das duas amostras. Presume-se então, que no instante $t_j$, ocorrem $d_j$ falhas, onde $n_j$ indivíduos estão submetidos há um risco no tempo inferior a $t_j$ na amostra combinada. Para as amostras $i$, $d_{ij}$ e $n_{ij}$ são, respectivamente, o número de falhas e indivíduos em risco imediatamente antes de $t_j$ com $i = 1, 2, \ldots, r$ e $j = 1, 2, \ldots, k$. Portanto, a diferença entre o número de falhas observadas e o número esperado em cada grupo fundamenta a estatística do teste.

Para dois grupos ($r = 2$), a estatística do teste *Logrank* é dada por:

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2}$$

em que:

- $d_{2j}$ é o número observado de falhas no grupo 2 no tempo $t_j$;
- $w_{2j}$ é o número esperado de falhas no grupo 2 no tempo $t_j$;
- $(V_j)_2$ é o elemento da matriz de variância correspondente ao grupo 2 no tempo $t_j$.

Para $r > 2$ grupo, a variância de $d_{ij}$ e a covariância são

$$(V_j)_{ii} = n_{ij}(n_j - n_{ij})d_j(n_j - d_j)n_j^{-2}(n_j - 1)^{-1}$$

e

$$(V_j)_{il} = -n_{ij}n_{li}d_j(n_j - d_j)n_j^{-2}(n_j - 1)^{-1}$$

respectivamente. Forma-se então a estatística $v$, somando todos os tempos distintos de falha:

$$v = \sum_{j=1}^{k} v_j,$$

em que $v$ é o vetor de dimensão $(r-1)\times 1$, com os elementos sendo as diferenças entre os totais observados e os esperados de falha. A função de sobrevivência pode ser expressa como:

$$T = v'V^{-1}v,$$

que, sob $H_0$ (igualdade das curvas), tem uma distribuição qui-quadrado com $r-1$ graus de liberdade para amostras grandes (Colosimo; Giolo, 2006, p. 59).

## 2.5 Modelos Paramétricos em Análise de Sobrevivência e Estimação dos Parâmetros

Além da estimação não paramétrica da função de sobrevivência é possível considerar as distribuições de probabilidade para a variável aleatória $T$ como tempo de falha. Nesta seção serão apresentadas alguns modelos, como exponencial, Weibull e log-normal.

### 2.5.1 Distribuição Exponencial

A distribuição exponencial é um dos modelos mais simples em relação ao tempo de falha. Sua função de densidade probabilística para um tempo de falha $T$ como variável aleatória é dada por:

$$f(t) = \frac{1}{\alpha}\exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \quad t \geq 0, \tag{2.5}$$

em que $\alpha > 0$ é o tempo médio de vida e $\alpha^2$, a variância. Assim, a função de sobrevivência $S(t)$ e a função de taxa de falha $\lambda(t)$ são expressas por:

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)\right\} \tag{2.6}$$

e

$$\lambda(t) = \frac{1}{\alpha} \quad \text{, para } t \geq 0. \tag{2.7}$$

Existe também, um modelo exponencial com dois parâmetros(Colosimo; Giolo, 2006). Nele tem-se um parâmetro $t_0$ que representa um tempo de período inicial onde a falha nunca

ocorre, conhecido como tempo de garantia. Para a função de densidade nesse caso, substitui-se $t$ por $t - t_0$, para $t_0 < t$, na expressão (2.5) .

### 2.5.2 Distribuição de Weibull

A distribuição de Weibull foi proposta por Weibull (1939). Nela a função de taxa de falha é monótona. Sua função de densidade de probabilidade, utilizando uma variável aleatória $T$ é dada por:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}, \quad t \geq 0, \tag{2.8}$$

em que o parâmetro de forma $\gamma$ e o de escala $\alpha$ são positivos. A unidade de medida de $t$ é a mesma apenas para o parâmetro $\alpha$, pois $\gamma$ não possui unidade.

Dessa forma, para a distribuição Weibull, a função de sobrevivência $S(t)$ e a de risco $\lambda(t)$ são expressas por:

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\} \tag{2.9}$$

e

$$\lambda(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}. \tag{2.10}$$

Conforme Colosimo e Giolo (2006), quando $\gamma = 1$, tem-se a distribuição exponencial, sendo esta um caso particular da distribuição de Weibull.

Para a distribuição de Weibull a média e a variância incluirão o uso da função gama definida por $\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx$, isto é

$$\mathbb{E}[T] = \alpha\,\Gamma(1 + 1/\gamma),$$

$$\text{Var}[T] = \alpha^2\left[\Gamma\left[1 + \left(\frac{2}{\gamma}\right)\right] - \Gamma\left[1 + \left(\frac{1}{\gamma}\right)\right]^2\right].$$

### 2.5.3 Distribuição Log-normal

A distribuição log-normal, como a distribuição de Weibull, é um modelo importante nas análises de sobrevivência, sendo muito utilizada nas áreas a da saúde e industriais. Sua função de densidade com uma variável aleatória $T$ é dada por:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\log t - \mu)^2\right\}, \quad t > 0,\ \sigma > 0, \tag{2.11}$$

em que $\mu$ é a média do logaritmo do tempo de falha e $\sigma$ é o desvio-padrão.

A distribuição log-normal possui uma relação com a normal, semelhante a de Weibull possui com a do valor extremo, o que facilita a apresentação e a análise de dados que decorrem da log-normal. "O logaritmo de uma variável com distribuição log-normal de parâmetros $\mu$ e $\sigma$ tem uma distribuição normal com média $\mu$ e desvio-padrão $\sigma$" (Colosimo; Giolo, 2006, p. 76).

Nesse contexto, o parâmetro $\mu$, que representa a escala na log-normal, corresponde à posição (ou locação) da distribuição normal. Já o parâmetro $\sigma$, que define a forma da log-normal, está relacionada à dispersão (ou escala) da distribuição normal.

Como as funções sobrevivência $S(t)$ e da taxa de falha $\lambda(t)$ da distribuição log-normal não possuem uma forma analítica explicitada, pode-se representá-las por:

$$S(t) = \Phi\left(\frac{-\log(t) + \mu}{\sigma}\right) \qquad \text{e} \qquad \lambda(t) = \frac{f(t)}{S(t)}$$

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

### 2.5.4 Método de Máxima Verossimilhança

O método da máxima verossimilhança é uma técnica usada para estimar os parâmetros de um modelo probabilístico, com base nos dados coletados. A ideia principal é considerar a função densidade que representa os dados e procurar os valores dos parâmetros que tornam os dados observados mais prováveis dentro desse modelo.

Por exemplo, se uma população de interesse em que todos indivíduos são não-censurados, a distribuição dos tempos de falha segue uma exponencial com parâmetro $\alpha$ (parâmetro a ser estimado), a função de densidade dada como na equação (2.5). A verossimilhança para uma

amostra de $n$ observações independentes $t_1, t_2, ..., t_n$ é definida como:

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

Essa função depende do parâmetro $\theta$, que pode representar um único valor ou um vetor de parâmetros, dependendo do modelo adotado. Em modelos como o log-normal, por exemplo, $\theta$ pode incluir média e desvio padrão $(\mu, \sigma)$. Determina-se assim o valor de $\theta$ que maximiza essa função de verossimilhança, ou seja, "encontrar o valor de $\theta$ que maximize a probabilidade da amostra observada ocorrer" (Colosimo; Giolo, 2006, p. 84).

Para observações não-censuradas, a função de verossimilhança $L(\theta)$ mostra contribuições que derivam da relação com a função de densidade, porém para observações censuradas a contribuição corresponde à função de sobrevivência $S(t)$, que representa a probabilidade do tempo de falha ser maior que o tempo censurado observado. Dessa forma, as observações podem ser separadas em dois grupos: que apresentam falhas e são as primeiras $r$ ordenadas, as não censuradas $(1, 2, ..., r)$ e aquelas $n - r$ seguintes que foram censuradas $(r + 1, r + 2, ..., n)$.

Neste caso de censura, temos:

$$L(\theta) = \prod_{i=1}^{r} f(t_i; \theta) \prod_{i=r+1}^{n} S(t_i; \theta), \tag{2.12}$$

ou reescrita como:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} [f(t_i; \theta)]^{\delta_i} [S(t_i; \theta)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [\lambda(t_i; \theta)]^{\delta_i} S(t_i; \theta), \end{aligned} \tag{2.13}$$

em que $\delta_i$ é a variável que indica a falha. O logaritmo da função de verossimilhança (2.12) e (2.13) são convenientemente utilizados para maximizar. Segundo Colosimo e Giolo (2006), os estimadores de máxima verossimilhança são os valores de $\theta$ que maximizam $L(\theta)$ ou equivalentemente o logaritmo de $L(\theta)$, isto é, $\log(L(\theta))$.

Dessa forma, pode-se encontrá-los com o seguinte sistema de equações:

$$U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = 0.$$

A seguir serão apresentados os desenvolvimentos do método de máxima verossimilhança para os modelos exponencial, Weibull e log-normal.

**Distribuição Exponencial**

Considere $n$ observações $(t_1, d_1), \ldots, (t_n, d_n)$ da variável $T$ com distribuição exponencial, dessa forma, a função de verossimilhança da distribuição exponencial é construída a partir da função de densidade de probabilidade em (2.5). Então:

$$\begin{aligned} L(\alpha) &= \prod_{i=1}^{n} \left[ \frac{1}{\alpha} \exp\left\{ -\left( \frac{t_i}{\alpha} \right) \right\} \right]^{\delta_i} \left[ \exp\left\{ -\left( \frac{t_i}{\alpha} \right) \right\} \right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[ \frac{1}{\alpha} \right]^{\delta_i} \exp\left\{ -\left( \frac{t_i}{\alpha} \right) \right\}. \end{aligned}$$

Calculando o logaritmo de $L(\alpha)$, obtem-se que:

$$\begin{aligned} \log(L(\alpha)) &= \sum_{i=1}^{n} \delta_i \log(1/\alpha) - \frac{1}{\alpha} \sum_{i=1}^{n} t_i \\ &= -\sum_{i=1}^{n} \delta_i \log(\alpha) - \frac{1}{\alpha} \sum_{i=1}^{n} t_i. \end{aligned}$$

Derivando $\log(L(\alpha))$ e igualando a zero a expressão resultante, para obter então, avaliando-a em $\alpha = \widehat{\alpha}$, o estimador de máxima verossimilhança para $\alpha$:

$$\frac{\partial \log L(\alpha)}{\partial \alpha} = -\frac{1}{\alpha^2} \sum_{i=1}^{n} \delta_i + \frac{1}{\alpha^2} \sum_{i=1}^{n} t_i.$$

Então,

$$\widehat{\alpha} = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} \delta_i} = \frac{\sum_{i=1}^{n} t_i}{r},$$

em que $\sum_{i=1}^{n} t_i$ é o tempo total sob teste. É importante ressaltar que se todas a observações fossem não censuradas, $\widehat{\alpha} = \bar{t}$, representa a média amostral.

**Distribuição de Weibull**

Considere $n$ observações $(t_1, \delta_1), \ldots, (t_n, \delta_n)$ da variável aleatória $T$ com distribuição Weibull, dessa forma, a função de verossimilhança da distribuição de Weibull é construída a partir da função de densidade de probabilidade em (2.8). Então:

$$\begin{aligned} L(\gamma, \alpha) &= \prod_{i=1}^{n} \left[ \frac{\gamma}{\alpha^{\gamma}} t_i^{\gamma-1} \exp\left\{ -\left(\frac{t_i}{\alpha}\right)^{\gamma} \right\} \right]^{\delta_i} \left[ \exp\left\{ -\left(\frac{t_i}{\alpha}\right)^{\gamma} \right\} \right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[ \frac{\gamma}{\alpha^{\gamma}} t_i^{\gamma-1} \right]^{\delta_i} \exp\left\{ -\left(\frac{t_i}{\alpha}\right)^{\gamma} \right\}. \end{aligned}$$

Calculando o logaritmo de $L(\gamma, \alpha)$, obtem-se que:

$$\begin{aligned} \log(L(\gamma, \alpha)) &= \log\left[ \prod_{i=1}^{n} \left( \frac{\gamma}{\alpha^{\gamma}} t_i^{\gamma-1} \right)^{\delta_i} \exp\left\{ -\left(\frac{t_i}{\alpha}\right)^{\gamma} \right\} \right] \\ &= \sum_{i=1}^{n} \delta_i \log(\gamma) - \sum_{i=1}^{n} \delta_i \gamma \log(\alpha) + (\gamma - 1) \sum_{i=1}^{n} \delta_i \log(t_i) - \alpha^{-\gamma} \sum_{i=1}^{n} t_i^{\gamma} \\ &= r \log(\gamma) - r\gamma \log(\alpha) + (\gamma - 1) \sum_{i=1}^{n} \delta_i \log(t_i) - \alpha^{-\gamma} \sum_{i=1}^{n} t_i^{\gamma}. \end{aligned}$$

Escrevendo $y_i = \log(t_i)$ obtem-se que:

$$\log(L(\mu, \sigma)) = -r \log(\sigma) + \sum_{i=1}^{n} \delta_i \left( \frac{y_i}{\sigma} \right) - \frac{r\mu}{\sigma} - \sum_{i=1}^{n} \exp\left\{ \frac{(y_i - \mu)}{\sigma} \right\},$$

que é mais simples que a função de verossimilhança da distribuição de Weibull logarítmica. Assim, fazendo o terceiro passo, no qual deriva-se essa última expressão com relação aos parâmetros $\mu$ e $\sigma$, igualando a zero e obtêm-se:

$$\begin{aligned} \frac{\partial L(\mu, \sigma)}{\partial \mu} &= \frac{1}{\widehat{\sigma}} \left[ -r + \sum_{i=1}^{n} \exp\left\{ \frac{(y_i - \widehat{\mu})}{\sigma} \right\} \right] = 0 \\ \frac{\partial L(\mu, \sigma)}{\partial \sigma} &= \frac{1}{\widehat{\sigma^2}} \left[ -r\widehat{\sigma} - \sum_{i=1}^{n} \delta_i y_i + r\widehat{\mu} + \sum_{i=1}^{n} \exp\left\{ \frac{(y_i - \widehat{\mu})}{\sigma} \right\} (y_i - \widehat{\mu}) \right] = 0 \end{aligned}$$

em que $L(\mu, \sigma) = \log(L(\mu, \sigma))$. A obtenção da solução desse sistema de equações demanda o uso de métodos numéricos tais como Newton Raphson.

**Distribuição Log-Normal**

Considere $n$ observações $(t_1, \delta_1), \ldots, (t_n, \delta_n)$ da variável aleatória $T$ com distribuição Log-Normal, tal que $Y_i = \log(T_i) \sim N(\mu, \sigma^2)$. Assim, a função de densidade de probabilidade de $T$ é dada por:

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

Então:

$$L(\mu, \sigma) = \prod_{i=1}^{n} [f(t_i; \mu, \sigma)]^{\delta_i} [S(t_i; \mu, \sigma)]^{1-\delta_i},$$

em que $S(t_i; \mu, \sigma)$ é a função de sobrevivência, dada por:

$$S(t_i; \mu, \sigma) = 1 - \Phi\left(\frac{\log t_i - \mu}{\sigma}\right),$$

sendo $\Phi(\cdot)$ a função de distribuição acumulada da normal padrão.

Substituindo a expressão de $f(t_i; \mu, \sigma)$ na equação da verossimilhança, obtém-se:

$$L(\mu, \sigma) = \prod_{i=1}^{n} \left[\frac{1}{t_i \sigma \sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\log t_i - \mu)^2\right\}\right]^{\delta_i} \left[1 - \Phi\left(\frac{\log t_i - \mu}{\sigma}\right)\right]^{1-\delta_i}.$$

Calculando o logaritmo natural da função de verossimilhança, obtém-se a log-verossimilhança:

$$\begin{aligned}\log L(\mu, \sigma) = &- \sum_{i=1}^{n} \delta_i \log(t_i) - r \log(\sigma) - \frac{r}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \delta_i (\log t_i - \mu)^2 \\ &+ \sum_{i=1}^{n} (1 - \delta_i) \log\left[1 - \Phi\left(\frac{\log t_i - \mu}{\sigma}\right)\right],\end{aligned}$$

em que $r = \sum_{i=1}^{n} \delta_i$ representa o número de observações não censuradas.

Por ultimo derivando a log-verossimilhança com relação aos parâmetros $\mu$ e $\sigma$ e igualando as expressões a zero, obtêm-se as equações de máxima verossimilhança:

$$\frac{\partial \log L(\mu,\sigma)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\delta_i(\log t_i - \mu) - \sum_{i=1}^{n}(1-\delta_i)\frac{\phi\left(\frac{\log t_i - \mu}{\sigma}\right)}{\sigma\left[1-\Phi\left(\frac{\log t_i - \mu}{\sigma}\right)\right]} = 0,$$

$$\frac{\partial \log L(\mu,\sigma)}{\partial \sigma} = -\frac{r}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}\delta_i(\log t_i - \mu)^2 - \sum_{i=1}^{n}(1-\delta_i)\frac{(\log t_i - \mu)\phi\left(\frac{\log t_i - \mu}{\sigma}\right)}{\sigma^2\left[1-\Phi\left(\frac{\log t_i - \mu}{\sigma}\right)\right]} = 0,$$

em que $\phi(\cdot)$ e $\Phi$ são a função densidade e a função acumulada da normal padrão, respectivamente.

A solução dessas equações fornece os estimadores de máxima verossimilhança $\widehat{\mu}$ e $\widehat{\sigma}$ e são encontradas a partir de métodos numéricos.

## 2.6 Modelos de Regressão para Dados de Sobrevivência

Em muitos estudos, tem-se o interesse de avaliar a relação de uma ou mais variáveis independentes com uma variável dependente. Nesta seção serão apresentados os modelos de regressão utilizados para dados censurados.

### 2.6.1 Modelo Linear para Dados de Sobrevivência

O modelo de regressão linear (Draper; Smith, 1998) é o mais conhecido e utilizado como ponto de partida. O modelo é dado por:

$$Y = \beta_0 + \beta_1 x + \epsilon, \tag{2.14}$$

em que $Y$ é a resposta (tempo até o evento), $x$ é a covariável, $\beta_0$ e $\beta_1$ são os parâmetros do modelo a serem estimados (intercepto e inclinação) e $\epsilon$ é o erro aleatório com distribuição normal que possui média 0 e variância constante $\sigma^2$.

O uso direto desse modelo não é adequado para a análise de sobrevivência, porque o tipo de resposta e o comportamento das variáveis geralmente não se ajustam bem. Portanto, tem-se duas formas para lidar deste problema:

1. Transformar as variáveis respostas;

2. Assumir uma distribuição assimétrica para os componentes estocásticos.

Essas duas abordagens conduzem à modelos equivalentes. O componente determinístico pode ser equivalente a um modelo linear com transformação logarítmica:

$$\exp\{\beta_0 + \beta_1 x\} \tag{2.15}$$

e para o erro utiliza-se a distribuição log-normal. Existem várias distribuições possíveis para o erro, além da log-normal, o que leva a diferentes modelos paramétricos, cada um com suas propriedades eque serão apresentados a seguir.

**Modelo de Regressão Exponencial**

Uma das distribuições assimétricas para o erro pode ser o modelo exponencial. Ao usar uma parte determinística junto com um erro que segue uma distribuição exponencial de média unitária ($f(\epsilon) = \exp\{-\epsilon\}$) para o erro obtem-se o seguinte modelo:

$$T = \exp\{\beta_0 + \beta_1 x\}\epsilon, \tag{2.16}$$

sendo este o modelo de regressão exponencial.

Ele possui uma relação não-linear entre $T$ e $x$ para o componente determinístico e para o erro uma distribuição assimétrica. "Na linguagem de modelos lineares generalizados (McCullagh; Nelder, 1989), tem-se uma função de ligação logarítmica e a resposta com distribuição exponencial"(Colosimo; Giolo, 2006, p.118-119).

O modelo (2.16) é não-linear em $T$ e pode ser transformado em linear aplicando o logaritmo de ambos os lados da equação para obter:

$$Y = \log(T) = \beta_0 + \beta_1 x + \nu, \tag{2.17}$$

em que $\nu = \log(\epsilon)$. A expressão (2.17) é semelhante a (2.14) mesmo com a distribuição dos erros sendo diferente com a (2.14), utilizando a normal. O erro $\nu$ segue uma distribuição chamada valor extremo padrão ($f(\nu) = \exp\{\nu - \exp\{\nu\}\}$), útil na análise de sobrevivência por leva diretamente ao modelo de riscos proporcionais.

A variável $x$ é linear em $Y$ em (2.17) e (2.16), dessa forma atua multiplicativamente em $T$. Para a função de sobrevivência de $Y$ para $y$, condicional a $x$ tem-se a expressão:

$$S(y \mid x) = \exp\{-\exp\{y - (\beta_0 + \beta_1 x)\}\}.$$

Agora para $T$ condicional a $x$ tem-se a expressão:

$$S(t \mid x) = \exp\left\{-\left(\frac{t}{\exp\{\beta_0 + \beta_1 x\}}\right)\right\}. \tag{2.18}$$

Depois de definir o modelo, o próximo passo é estimar os parâmetros. No caso específico do modelo (2.17) é preciso fazer inferência sobre o vetor de parâmetros $\boldsymbol{\theta} = (\beta_0, \beta_1)$, utilizando por exemplo o método de máxima verossimilhança.

Para a construção da função de verossimilhança, precisa-se separar dois grupos: os que sofreram falhas e os que foram censurados. Nesse caso, a contribuição do indivíduo com as covariáveis $x$ para a função de verossimilhança é a chance de que ele tenha sobrevivido até aquele momento, ou seja, a probabilidade de que o tempo até a falha seja maior que $t_i$ dada pela função de sobrevivência $S(t_i \mid x_i)$. Assumindo que os dados são independentes entre si, a função de verossimilhança para o modelo representado em (2.17) e uma amostra de tamanho $n$ com $(t_i, d_i, x_i)$ para $i = 1, \ldots, n$ é dado por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[f(y_i, \mid x_i)^{\delta_i}\right] \left[S(y_i, \mid x_i)^{(1-\delta_i)}\right], \tag{2.19}$$

em que $y_i = \log(t_i)$, ou para os modelos que possuem a forma de (2.16) dados por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[f(t_i, \mid x_i)^{\delta_i}\right] \left[S(t_i, \mid x_i)^{(1-\delta_i)}\right]. \tag{2.20}$$

Substituindo as funções de densidade e sobrevivência, para obter os estimadores de máxima verossimilhança, pelas distribuições do valor extremo em (2.19) ou da exponencial em

(2.20) tem-se que:

$$\begin{aligned}\log L(\boldsymbol{\theta}) &= \sum_{i=1}^{n} [\delta_i \left(y_i - \beta_0 - \beta_1 x_i - \exp\left\{y_i - \beta_0 - \beta_1 x_i\right\}\right) \\ &\quad -(1-\delta_i)\exp\left\{y_i - \beta_0 - \beta_1 x_i\right\}] \\ &= \sum_{i=1}^{n} \left[\delta_i (y_i - \beta_0 - \beta_1 x_i) - \exp\left\{y_i - \beta_0 - \beta_1 x_i\right\}\right]. \end{aligned} \tag{2.21}$$

Os valores de $\boldsymbol{\theta} = (\beta_0, \beta_1)$ são os estimadores de máxima verossimilhança feita ao maximizar a função log-verossimilhança. Para isso, derivamos essa função em relação a cada parâmetro e igualamos a zero, formando um sistema de equações, segundo Colosimo e Giolo (2006) e cuja solução é obtida por métodos numéricos.

**Modelo de Regressão Weibull**

Para a introdução de modelagem com dados de sobrevivência, o modelo exponencial é a maneira mais simples, porém não engloba a maioria das situações práticas. Por exemplo, segundo Nelson (1990), apenas cerca de 10% dos produtos industriais têm tempos de vida que seguem a distribuição exponencial. Uma maneira de tornar o modelo exponencial mais generalizado é adicionando um parâmetro extra de escala, que em modelos lineares, assume-se que os erros são normalmente distribuídos com média zero e variância constante $\sigma^2$. Essa variância pode assumir qualquer valor positivo, diferentemente da normal padrão, cuja variância é igual a 1. O modelo (2.17) é expresso na forma $Y = \log(T) = \beta_0 + \beta_1 x + \sigma\nu$ ou havendo a presença de covariáveis como:

$$Y = \log(T) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \sigma\nu = \mathbf{x}'\boldsymbol{\beta} + \sigma\nu, \tag{2.22}$$

com $\mathbf{x}' = (1, x_1, \ldots, x_p)$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$. Observa-se que na exponencial, considera $p = 1$. "Este modelo é conhecido como modelo de regressão Weibull, pois $T$ deve ter uma distribuição de Weibull para que $\log(T)$ tenha uma distribuição do valor extremo com parâmetros de escala $\sigma$"(Colosimo; Giolo, 2006, p. 121). Portanto, a função de sobrevivência de $Y$ condicional a $\mathbf{x}$ é dada por:

$$S(y \mid \mathbf{x}) = \exp\left\{-\exp\left\{\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right\}\right\}$$

e com $T$ condicional a $\mathbf{x}$ tem-se que:

$$S(t \mid \mathbf{x}) = \exp\left\{-\left(\frac{t}{\exp(\mathbf{x}'\boldsymbol{\beta})}\right)^{1/\sigma}\right\}.$$

A função de verossimilhança é obtida por meio da equação (2.20) e detalhes da estimação de $\beta$ podem ser obtidos em Colosimo e Giolo (2006).

**Modelo de Regressão Log-Normal**

Para o modelo de regressão Log-Normal assume-se que o logaritmo do tempo de sobrevivência segue uma distribuição normal, ou seja, $\log(T) \sim N(\mu, \sigma^2)$. O parâmetro $\sigma$ atua como um parâmetro de escala, controlando a dispersão dos tempos de sobrevivência no domínio logarítmico. Assim, como nos modelos lineares, considera-se que os erros são normalmente distribuídos com média zero e variância constante $\sigma^2$. Essa variância pode assumir qualquer valor positivo, representando a variabilidade inerente dos tempos de sobrevivência em relação ao valor esperado, diferentemente da normal padrão, cuja variância é igual a 1. O modelo é dado por:

$$Y = \log(T) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \sigma\nu = \mathbf{x}'\boldsymbol{\beta} + \sigma\nu, \tag{2.23}$$

onde

$$\mu(\mathbf{x}) = (\mathbf{x}'\boldsymbol{\beta})$$

com $\mathbf{x}' = (1, x_1, \ldots, x_p)$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$.

Assim, a função de sobrevivência de $Y$ condicional a $\mathbf{x}$ é dada por:

$$S(y \mid \mathbf{x}) = 1 - \Phi\left(\frac{y - \mu(\mathbf{x})}{\sigma}\right)$$

em que $\Phi(\cdot)$ representa a função de distribuição acumulada da Normal Padrão e assim, a função densidade condicional de $T$ pode ser expressa como:

$$f(t \mid \mathbf{x}) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(t) - \mu(\mathbf{x})}{\sigma}\right)^2\right], \quad t > 0.$$

Da mesma forma que o modelo de regressão Weibull, mais detalhes sobre a função por máxima verossimilhança podem ser encontradas em Klein e Moeschberger (2003).

## 2.7 Adequação do Modelo Ajustado

Na análise dos dados, avaliar se o modelo ajustado está realmente bom é uma etapa importante. Para isso, no modelo de regressão linear usual, realiza-se de uma análise gráfica dos resíduos. Por meios desses resíduos, pode-se avaliar a distribuição dos erros, sendo utilizada principalmente para ponderar e rejeitar modelos inapropriados. Nesses casos, não seria de utilidade considerar que um modelo paramétrico esta adequado para a análise.

Na área de Análise de Sobrevivência os resíduos de Cox-Snell Cox e Snell (1968) ajudam a avaliar se o modelo está bem ajustado aos dados. Eles podem ser determinados por:

$$\widehat{e_i} = \widehat{\Lambda}(t_i \mid \mathbf{x}_i), \tag{2.24}$$

onde $\widehat{\Lambda}(\cdot)$ é a função de risco acumulado obtida por meio do modelo ajustado para $i = 1, \ldots, n$. Para diferentes modelos, tem-se(Colosimo; Giolo, 2006):

1. Modelo de regressão exponencial:

$$\widehat{e_i} = \left[ t_i \exp\left\{ -\mathbf{x}_i' \widehat{\boldsymbol{\beta}} \right\} \right].$$

2. Modelo de regressão Weibull:

$$\widehat{e_i} = \left[ t_i \exp\left\{ -\mathbf{x}_i' \widehat{\boldsymbol{\beta}} \right\} \right]^{\widehat{\gamma}}.$$

3. Modelo de regressão log-normal:

$$\widehat{e_i} = -\log\left[ 1 - \Phi\left( \frac{\log(t_i) - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}}{\widehat{\sigma}} \right) \right].$$

Esses resíduos devem seguir uma distribuição exponencial padrão se o modelo estiver correto. Para verificar isso, pode-se construir gráficos comparando $\widehat{e_i}$ com sua função de risco acumulado estimada. "Assim, o gráfico $\widehat{e_i}$ *versus* $\widehat{\Lambda}(\widehat{e_i})$ deve ser aproximadamente uma reta com inclinação 1, quando o modelo exponencial for adequado, uma vez que $\widehat{\Lambda}(\widehat{e_i}) = -\log(\widehat{S}(\widehat{e_i}))$"(Colosimo; Giolo, 2006, p. 124). A curva de sobrevivência dos resíduos, ajudam a verificar a capacidade do modelo ajustar-se à função de sobrevivência empírica.

Segundo Lawless (1982), quando há poucos dados censurados e o modelo é exponencial ou Weibull, é possível considerar os resíduos censurados como se fossem completos, e pode-se defini-los por:

$$\widehat{e_i} = \left[ t_i \exp(-\mathbf{x}_i'\widehat{\boldsymbol{\beta}}) \right]^{\widehat{\gamma}} + 1 = \widehat{\Lambda}(t_i \mid \mathbf{x}_i) + 1.$$

Os resíduos de Cox–Snell não permitem identificar qual é o problema específico quando o gráfico apresenta comportamento não linear, porém eles ainda são úteis para analisar o ajuste global do modelo.

## 2.8 Critério de Informação de Akaike (AIC)

O Critério de Informação de Akaike (AIC) foi proposto por Akaike (1973) como uma extensão do princípio da máxima verossimilhança, fundamentado na teoria da informação. O autor buscava uma forma de selecionar modelos estatísticos que representassem adequadamente o processo gerador dos dados, equilibrando a qualidade do ajuste e a complexidade do modelo. Em termos conceituais, o AIC estima a quantidade de informação perdida ao se aproximar o modelo verdadeiro por um modelo candidato.

De acordo com a Elsevier (s.d.), o AIC mede esse equilíbrio penalizando a complexidade do modelo (número de parâmetros), enquanto recompensa o bom ajuste aos dados observados. Assim, valores menores de AIC indicam modelos mais adequados, pois sugerem menor perda de informação e maior parcimônia. Matematicamente, o AIC é dado por:

$$\text{AIC} = 2k - 2\ln(\widehat{L}),$$

em que $k$ representa o número de parâmetros estimados e $\hat{L}$ corresponde ao valor máximo da função de verossimilhança. O termo $2k$ atua como penalização pela complexidade do modelo, enquanto $-2\ln(\hat{L})$ reflete o grau de ajuste aos dados. O modelo com o menor valor de AIC é, portanto, considerado o mais informativo entre os candidatos analisados.

# 3. Material e Métodos

Este capítulo baseou-se na revisão teórica dos conceitos de análise de sobrevivência, suas principais funções probabilísticas e os modelos estatísticos utilizados para descrever o tempo até a ocorrência de um evento apresentados por Colosimo e Giolo (2006) e aplicação dos conceitos por meio de simulações no programa estatístico R.

Em situações com dados reais, a estimação dos parâmetros desses modelos é realizada por meio do método da máxima verossimilhança, que busca encontrar os valores que maximizam a probabilidade de observar os dados disponíveis e após isso a adequação do modelo é analisada pelos resíduos. Agora no caso dos dados simulados, a estimação dos parâmetros veio por meio da criação de um cenário hipotético, para assim ajustar os modelos com a máxima verossimilhança, compará-los utilizando a log-verossimilhança e concluir qual modelo mais se encaixa a situação criada.

## 3.1 Simulação no RStudio

### 3.1.1 Simulação dos Modelos Probabilísticos

Utilizou-se a linguagem R por meio do RStudio para a execução da análise por simulação junto com o pacote `survival`. A simulação iniciou-se pela definição de uma semente "set.seed", seguindo de uma população composta de n indivíduos e as funções "rexp" (Exponencial), "rweibul" (Weibull) e "rlnorm" (Log-Normal), explicitadas no programa, respectivamente, como rexp(n, rate = 0.05), rweibull(n, shape = 2, scale = 10) e rlnorm(n, meanlog = 2, sdlog = 0.5).

A censura aleatória foi aplicada com probabilidade 50% para cada indivíduo (sample(c(0,1), n, prob = c(0.5, 0.5), replace = TRUE)), sendo 1 indicador de evento observado e 0 censura. Curvas de sobrevida foram estimadas por Kaplan-Meier ("survfit") e os modelos foram ajustados utilizando o método de máxima verossimilhança com "survreg", obtendo-se os parâmetros de escala e forma conforme cada distribuição. Os parâmetros estimados foram $\alpha$ para a exponencial, o parâmetro de escala $\alpha$ e forma $\gamma$ para Weibull, junto com $\mu$ e $\sigma$ para a

log-normal.

As funções de sobrevivência ajustadas conforme os modelos foram comparadas visualmente às curvas de Kaplan-Meier e, adicionalmente, o valor da log-verossimilhança final de cada modelo foi adquirido para avaliação da adequação dos modelos aos dados simulados dos ajustes.

### 3.1.2 Simulação do conjunto de Dados com Variáveis Independentes

A simulação iniciou-se pela definição de uma semente "set.seed", assim como na primeira análise, seguindo de uma população composta de *n* indivíduos, distribuídos aleatoriamente entre os sexos masculino e feminino ("sample"), com a idade de cada indivíduo sendo gerada a partir de uma distribuição normal ("rnorm"). O tempo limite do experimento `tempo_final` foi limitado a no máximo 12 meses e arredondado para uma casa decimal para facilitar a análise. Para o tempo de sobrevivência inicial `tempo_base` utilizou-se a distribuição de Weibull, da qual pode-se garantir parâmetros diferentes para homens e mulheres (variações associadas ao sexo).

Em seguida, esse tempo foi ajustado conforme a idade de cada participante e reduzindo 0,1 unidade para cada ano acima de 60 anos, porém em casos nos quais o valor resultante ficou abaixo de 0,1 mês, atribuiu-se um tempo aleatório entre 0,1 e 0,5 mês, garantindo que todos tivessem um acompanhamento positivo. Para indicar o desfecho, adotou-se o valor “1” para os casos em que ocorreu o evento (remissão) e “0” quando houve censura (morte ou evasão), considerando 80% de probabilidade de evento para tempos inferiores ao limite máximo de acompanhamento. Estimou-se a curva de sobrevida por sexo por meio do estimador de Kaplan-Meier, comparando-se as curvas entre os sexos, com o teste *logrank* ($\rho = 0$).

# 4. Resultados e Discussão

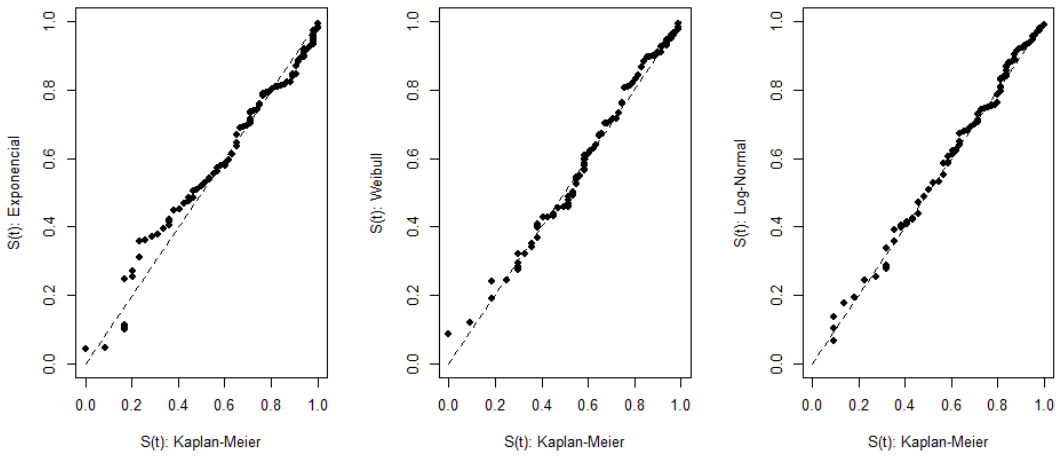
Neste capítulo serão apresentados os resultados de dois conjuntos de dados diferentes, junto com a discussão das análises feitas por meio dos ajustes dos modelos probabilísticos.

## 4.1 Simulação dos modelos probabilísticos

Considerando um estudo que acompanhou um grupo de pacientes com câncer de mama submetidos à cirurgia e ao tratamento quimioterápico, foram simulados 100 indivíduos com o tempo até a remissão total. O código utilizado para as simulações encontram-se no Anexo 1. Ajustes foram realizados utilizando o método da máxima verossimilhança.

Na Figura 4.1 tem-se gráficos de concordância dos modelos exponencial, Weibull e log-normal, respectivamente. No eixo X tem-se o método de Kaplan-Meier com um $S(t)$ estimado empiricamente e no eixo Y as funções de sobrevivência dos modelos citados. Dessa forma, é possível avaliar a concordância ponto a ponto entre Kaplan-Meier e os modelos trazendo uma avaliação rápida e global dos modelos. Quanto mais afastado os pontos, menos adequado é o modelo, pela análise gráfica.

**Figura 4.1:** Gráficos das sobrevivências estimadas por Kaplan-Meier *versus* as estimada pelos modelos exponencial, Weibull e log-normal, respectivamente.

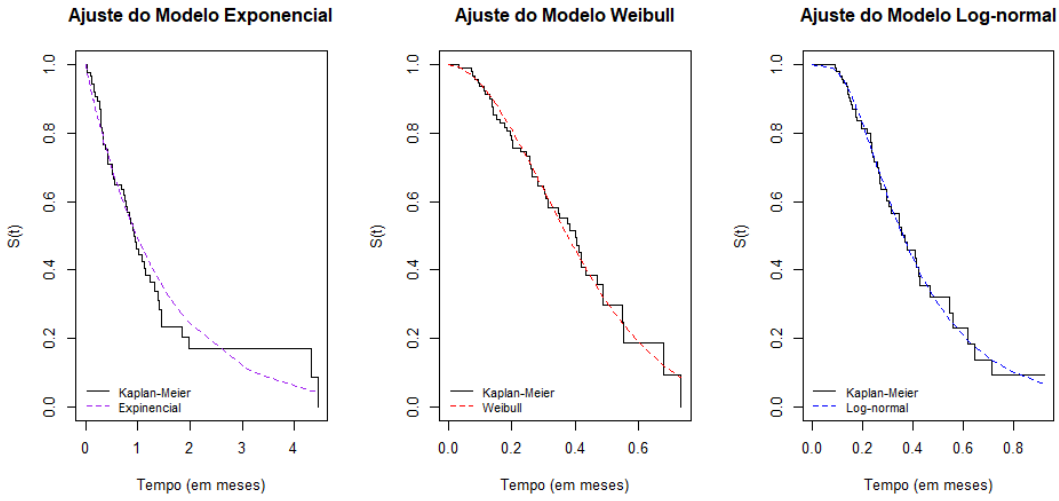



Fonte: Própria autoria (2025).

Portanto analisando os resultados, o modelo exponencial aparentemente não possui um bom ajuste. A curva apresenta-se afastada da reta y = x, mais do que nos casos do modelo de Weibull e log-normal.

Agora utilizando-se dos ajustes dos modelos por meio da máxima verossimilhança, obtêm-se gráficos nos quais é possível interpretar as tendências temporais. Na Figura 4.2, observam-se esses gráficos com os dados estimando as curvas de sobrevivência pelos modelos anteriormente citados, junto da curva estimada por Kaplan-Meier empiricamente. Dessa forma, é possível acompanhar como o modelo demonstra a tendência de sobrevivência. Nas regiões onde ocorrem desvios, menos adequado será o ajuste em intervalos específicos de tempo. De modo geral a construção destes gráficos permite avaliar visualmente como as curvas dos modelos, acompanha os dados simulados ao longo do tempo.

**Figura 4.2:** Gráfico de sobrevivência *versus* o tempo em meses por Kaplan-Meier e os modelos paramétricos.




Fonte: Própria autoria (2025).

Portanto analisando os resultados graficamente, o modelo exponencial neste contexto também não possui bom ajuste de acordo com os dados.

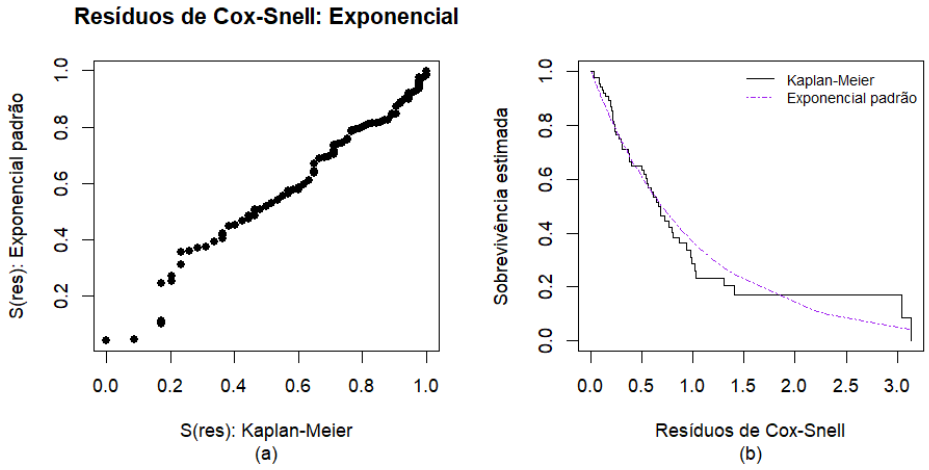
Para adquirir resultados quantitativos de qual o melhor ajuste, foi utilizada a log-verossimilhança que corroborou nessas observações, pois o modelo log-Normal apresentou a maior log-verossimilhança de -5,1411, seguido pelo Weibull de -9,8428 e, por último, o exponencial de -70,9978, indicando que o log-Normal forneceu o ajuste mais próximo e que a exponencial permanece o ajuste menos favorável.

Os gráficos presentes na Figura 4.2 são a base para calcular os resíduos de Cox-Snell, sendo as diferenças dos dados simulados e dos dados obtidos por meio dos modelos ajustados. Dessa forma nas Figuras 4.3, 4.4 e 4.5 observam-se gráficos sobre os resíduos de Cox-Snell para os ajustes exponencial, Weibull e log-normal onde tem-se dois gráficos.

O gráfico (a), possui um eixo X que representa a função de sobrevivência dos resíduos, calculada pelo estimador de Kaplan-Meier e um eixo Y que representa a função de sobrevivência dos mesmos resíduos, mas assumindo uma distribuição referente ao modelo exponencial, Weibull e log-normal. Para um bom ajuste os pontos deveriam se alinhar perfeitamente em uma linha diagonal de 45 graus.

Já no gráfico (b), mostra-se uma comparação entre a função de sobrevivência dos resíduos, estimada de forma não-paramétrica e a exponencial, Weibull e log-normal padrão, que mostram a função de sobrevivência ajustada dos resíduos de Cox-Snell, assim quanto mais alinhados estiverem as funções, melhor ajustado o modelo será.

**Figura 4.3:** Análise dos resíduos de Cox-Snell do modelo de regressão exponencial ajustado.

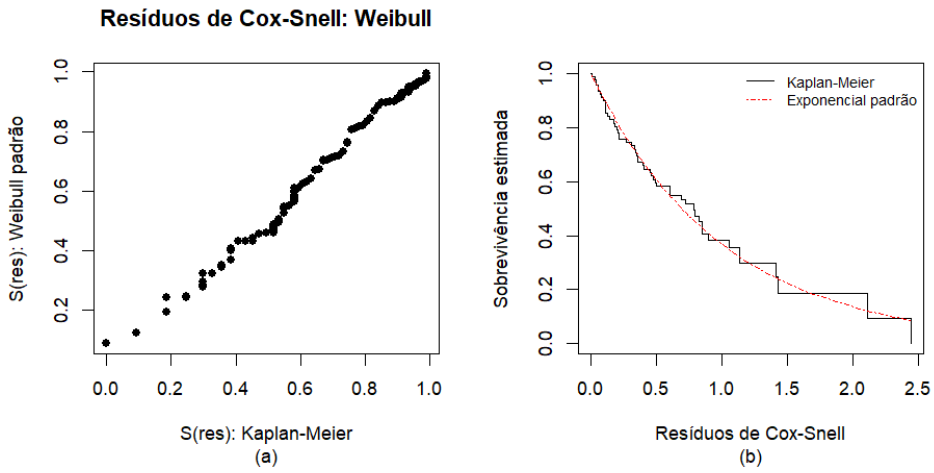



Fonte: Própria autoria (2025).

Com essa análise gráfica dos resíduos, o modelo exponencial não apresenta bom ajustes para os dados simulados.

Se o modelo de sobrevivência Weibull for adequado aos dados, os resíduos de Cox-Snell deverão se comportar de acordo com essa distribuição exponencial padrão.

**Figura 4.4:** Análise dos resíduos de Cox-Snell do modelo de regressão Weibull ajustado.

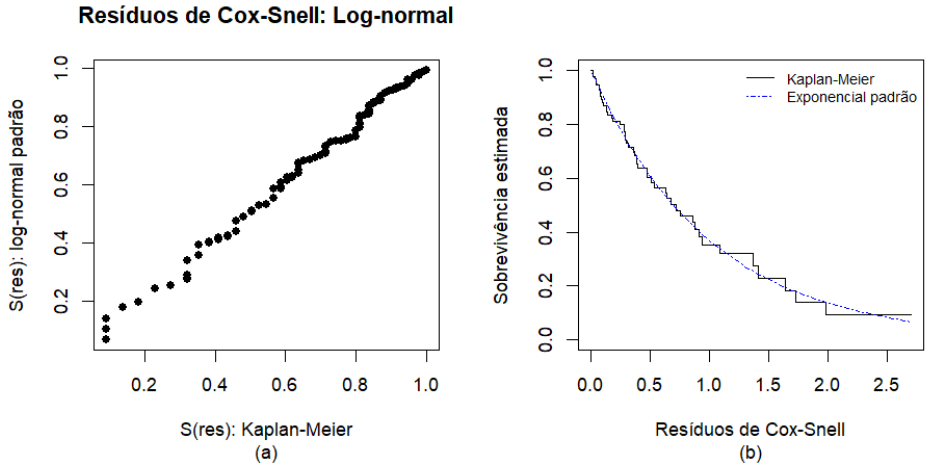



Fonte: Própria autoria (2025).

Dessa forma, as análises gráficas dos resíduos mostra que o modelo de Weibull possui um bom ajuste para os dados.

Para os resíduos de Cox-Snell da log-normal, a função de sobrevivência ajustada dos resíduos será adequada quando puder transformar qualquer função de sobrevivência em exponencial padrão, a partir da log-normal padrão.

**Figura 4.5:** Análise dos resíduos de Cox-Snell do modelo de regressão log-normal.




Fonte: Própria autoria (2025).

Portanto de todas as análises gráficas dos resíduos expõem que o modelo de regressão log-normal, possui um bom ajuste , gerado dos modelos respectivos.

Na Tabela 4.1 há a comparação quantitativa dos parâmetros simulados e os estimados por máxima verossimilhança, sendo essa uma maneira de mostra quão próximos os valores estimados estão dos valores verdadeiros, ou seja, apenas a precisão da estimação de cada parâmetro individual. Ela não avalia diretamente a qualidade do ajuste do modelo ao comportamento dos dados ao longo do tempo.

**Tabela 4.1:** Comparação entre os valores verdadeiros e estimados dos parâmetros dos modelos apresentados.

| Modelo | Parâmetro | Verdadeiro | Estimado | Erro Absoluto | Erro Relativo (%) |
|---|---|---|---|---|---|
| Exponencial | $\alpha$ | 0.667 | 1.441 | 0.774 | 116.15 |
| Weibull | $\alpha$ | 0.333 | 0.465 | 0.132 | 39.65 |
| | $\gamma$ | 2.000 | 1.881 | 0.119 | 5.95 |
| Log-normal | $\mu$ | -1.401 | -0.998 | 0.402 | 28.71 |
| | $\sigma$ | 0.500 | 0.626 | 0.126 | 25.27 |

Fonte: Própria autoria (2025).

O cálculo do erro relativo é importante para comparar ajustes entre parâmetros de escalas diferentes, já que o erro absoluto sozinho pode ser enganoso.

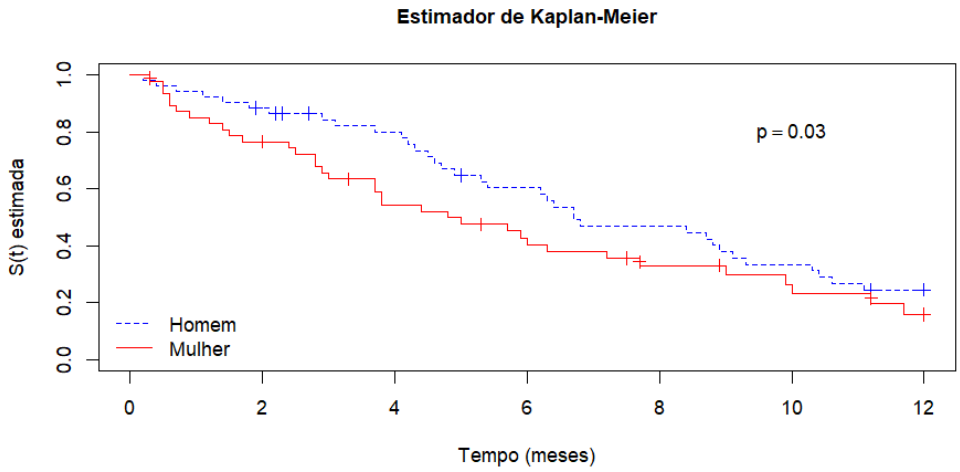
## 4.2 Simulação do conjunto de Dados com Variáveis Independentes

Considerando um estudo que acompanhou um grupo de pacientes com um certo tipo de câncer pulmonar, a taxa de remissão foi feita com base nas variáveis gênero e idade, assim foram simulados 100 pacientes com idade entre 30 e 60 anos com o tempo até a remissão total em meses. Os modelos considerados foram $Y = \log(T) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ em que $x_1$ refere-se ao gênero e $x_2$ a idade em anos. Os comandos encontram-se no Anexo 2. Foram obtidas duas curvas pela estimação de Kaplan-Meier para fazer a comparação e as simulações dos dados vieram da distribuição Normal e Weibull e obter os resultados.

Na Figura 4.6, há um gráfico no qual nota-se a relação entre o tempo em meses até o evento (remissão) ou a censura (morte/evasão) e a probabilidade estimada de não ter ocorrido o evento até para cada tempo. Assim, observa-se que para as mulheres, a remissão leva menos

tempo para ocorrer em comparação com os homens, pois a curva delas decai mais rapidamente. Por fim, as curvas são estatisticamente significativas ao nível de 5% (teste logrank), sugerindo associação entre o sexo e a sobrevida do paciente.

**Figura 4.6:** Estimador de Kaplan-Meier com diferenciação da remissão em homens e mulheres e valor de p = 0.03, dado pela análise com teste "logrank".




Fonte: Própria autoria (2025).

Agora ajustando os modelos de regressão paramétricos exponencial, Weibull e log-normal, pode-se quantificar as diferenças entre os grupos, ver qual modelo descreve melhor os dados e interpretar o efeito das variáveis independentes sobre o tempo até o evento.

Para a Figura 4.7 utilizou-se um exemplo com dados simulados que facilitam a visualização gráfica da probabilidade de remissão ao longo do tempo para homens e mulheres de 60 anos, usando os modelos utilizados no ajuste para tentar aproximar a curva real de Kaplan-Meier.

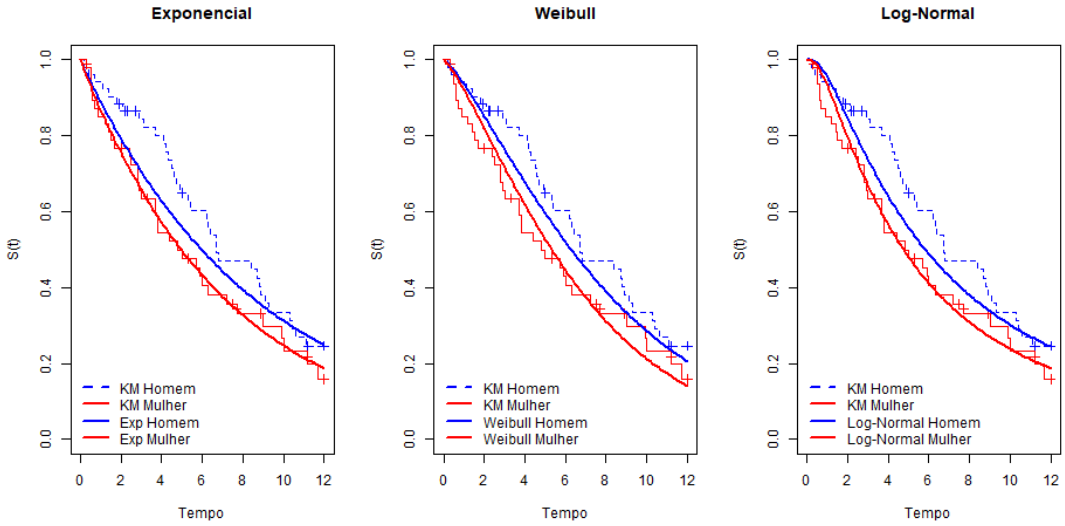
Na distribuição exponencial, observa-se que as curvas decrescem de forma aproximadamente constante ao longo do tempo, refletindo a suposição de uma taxa de risco constante. No entanto, esse modelo não representa bem a diferença entre os sexos, pois tende a superestimar a sobrevivência dos homens e subestimar a das mulheres. Ainda assim, nota-se que, para as mulheres, a curva exponencial apresenta maior proximidade em relação à estimativa de Kaplan-Meier.

No caso da Weibull, há uma melhor adequação geral às curvas empíricas de Kaplan-Meier tanto para homens, quanto para mulheres. O formato flexível desse modelo permite capturar a variação da taxa de risco ao longo do tempo, o que se reflete na boa concordância

visual, especialmente entre as mulheres, cuja curva paramétrica se ajusta de forma mais próxima à observada.

Já a distribuição log-normal reproduz bem o comportamento inicial das curvas, acompanhando a queda mais acentuada da sobrevivência (ou a remissão mais rápida) nos primeiros períodos. Entretanto, tende a suavizar o comportamento no final do seguimento. Nesse modelo, novamente, a curva ajustada para as mulheres apresenta maior similaridade com a de Kaplan-Meier, indicando melhor adequação em relação ao grupo masculino.

**Figura 4.7:** Estimador de Kaplan-Meier com diferenciação da remissão em homens e mulheres e valor de p = 0.03, dado pela análise com teste "logrank".




Fonte: Própria autoria (2025).

Se a curva do modelo paramétrico estiver próxima da curva de Kaplan-Meier, significa que o modelo descreve bem a distribuição dos tempos até o evento. Utilizando a comparação dos modelos com AIC tem-se que: exponencial possui grau de liberdade 3 e AIC = 433,3928; Weibull grau de liberdade 4 e AIC = 429,9131 ; log-normal grau de liberdade 4 e AIC = 429,7107. Portanto a log-normal, tanto graficamente, quanto pela comparação quantitativa dos modelos, possui melhor ajuste para a situação de exemplo para homens e mulheres de 60 anos.

### 4.2.1 Coeficientes dos modelos paramétricos

Nas Tabelas 4.2, 4.3 e 4.4 apresentam-se os coeficientes estimados para os modelos exponencial, Weibull e log-normal. Esses coeficientes foram obtidos a partir do ajuste dos

modelos paramétricos aos dados simulados. O parâmetro $\beta_0$ representa a proporcionalidade entre as curvas de sobrevivência, tendo os homens como categoria de referência ($\beta_1 = 0$). O coeficiente $\beta_2$, associado à idade, apresenta valores reduzidos e semelhantes entre os grupos, sugerindo ausência de interação entre idade e sexo.

As mulheres possuem valores menores, pois o $\beta_1$ apresenta valores negativos à elas e portanto a velocidade com que se alcança a remissão para mulheres é mais rápida. O $\beta_2$ apresenta coeficientes negativos para as idades no modelo de regressão exponencial e Weibull, o que indica que quanto maior a idade do paciente, menor será o tempo até a remissão. Dessa forma em pacientes acima de 60 anos possuem as chances de remissão diminuídas conforme mais velhos vão ficando. Os valores de p para os $\beta_0$ são estatisticamente significativos, porém para os outros valores, o sexo e idade não afetam o desfecho de forma significativa.

**Tabela 4.2:** Coeficientes estimados ($\beta_0$, ($\beta_1$ e $\beta_2$) para o modelo exponencial e valor de p.

| **Variável** | **Estimativa** | **Erro Padrão** | **p-valor** |
|---|---|---|---|
| $\beta_0$ | 2.363 | 0.662 | 0.0003 |
| $\beta_1$ | -0.290 | 0.241 | 0.2292 |
| $\beta_2$ | -0.002 | 0.011 | 0.8706 |

Fonte: Própria autoria (2025)

**Tabela 4.3:** Coeficientes estimados ($\beta_0$, ($\beta_1$ e $\beta_2$) para o modelo Weibull e valor de p.

| **Parâmetro** | **Estimado** | **Erro Padrão** | **p-valor** |
|---|---|---|---|
| $\beta_0$ | 2.359 | 0.457 | 2.5e-07 |
| $\beta_1$ | -0.217 | 0.167 | 0.1934 |
| $\beta_2$ | -0.003 | 0.007 | 0.7134 |
| Escala | -0.369 | 0.098 | 0.0002 |

Fonte: Própria autoria (2025)

**Tabela 4.4:** Coeficientes estimados ($\beta_0$, ($\beta_1$ e $\beta_2$) para o modelo log-normal e valor de p

| **Parâmetro** | **Estimado** | **Erro Padrão** | **p-valor** |
|---|---|---|---|
| $\beta_0$ | 1.593 | 0.581 | 0.0061 |
| $\beta_1$ | -0.324 | 0.211 | 0.1252 |
| $\beta_2$ | 0.005 | 0.009 | 0.5880 |
| Escala | -0.019 | 0.086 | 0.8169 |

Fonte: Própria autoria (2025)

# 5. Conclusão

O presente trabalho teve como propósito compreender e aplicar os modelos de regressão paramétricos em dados de sobrevivência, integrando a teoria estatística com experimentos simulados. A partir da revisão conceitual, foi possível consolidar o entendimento sobre as funções fundamentais da análise de sobrevivência como a função de sobrevivência, taxa de falha, função de risco acumulado e o papel das observações censuradas, que são essenciais para a modelagem de tempos até o evento.

Nas simulações realizadas no RStudio, com o uso do pacote `survival`, observou-se que os modelos exponencial, Weibull e log-normal produzem estimativas coerentes com as propriedades teóricas das respectivas distribuições. A comparação entre as curvas estimadas pelos modelos e o estimador não paramétrico de Kaplan-Meier demonstrou que todos apresentaram boa aproximação visual, porém com diferenças sutis de ajuste dependendo da forma dos dados.

O modelo exponencial, por assumir taxa de falha constante, foi o menos flexível, sendo adequado apenas para fenômenos com risco uniforme ao longo do tempo. Já o modelo de Weibull mostrou-se o mais abrangente e eficiente, pois permite modelar situações com taxas de falha crescentes, decrescentes ou constantes, incluindo a distribuição exponencial como caso particular. O modelo log-normal, por sua vez, apresentou boa performance para cenários em que o risco de falha aumenta até um ponto máximo e depois diminui — comportamento típico em estudos clínicos e de durabilidade biológica.

A análise dos resíduos de Cox-Snell confirmou a adequação dos ajustes, indicando boa consistência para os modelos Weibull e log-normal, o que reforça a importância de validar estatisticamente o modelo selecionado.

De modo geral, os resultados obtidos evidenciam que os modelos paramétricos de regressão em análise de sobrevivência são ferramentas versáteis e eficazes na investigação de fenômenos associados ao tempo até um evento, permitindo inferências mais precisas e interpretações consistentes sobre os fatores que influenciam a sobrevivência.

# Referências


AKAIKE, Hirotugu. *Information Theory and an Extension of the Maximum Likelihood Principle*. Edição: B. N. Petrov e F. Csáki. **Proceedings of the Second International Symposium on Information Theory**, Akadémiai Kiadó, Budapest, p. 267–281, 1973.

BEWICK, Viv; CHEEK, Liz; BALL, Jonathan. *Statistics review 12: Survival analysis*. **Critical Care**, v. 8, n. 5, p. 389–394, 2004. DOI: 10.1186/cc2955. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC1065034/.

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

COUCHOUD, Cecile *et al. Restricted mean survival time over 15 years for patients starting renal replacement therapy*. **Kidney International**, v. 91, n. 6, p. 1347–1355, 2017. DOI: 10.1016/j.kint.2016.12.013. Disponível em: https://pubmed.ncbi.nlm.nih.gov/28057870/.

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

DRAPER, Norman R.; SMITH, Harry. *Applied regression analysis*. New York: John Wiley & Sons, 1998. v. 326.

ELSEVIER. *Akaike Information Criterion – An Overview*. [*S. l.: s. n.*]. https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/akaike-information-criterion. Accessed: 29 Oct. 2025.

HANDBOOK, NIST Engineering Statistics. *1.3.6.6.9. Lognormal Distribution*. [*S. l.: s. n.*], 2025. https://www.itl.nist.gov/div898/handbook/eda/section3/eda3669.htm. Accessed: 2025-10-28.

IHAKA, Ross; GENTLEMAN, Robert. *The R Project: A Brief History and Thoughts About the Future*. [*S. l.: s. n.*], 1993. Disponível no site oficial do R. Disponível em: https://www.r-project.org/.

KAPLAN, Edward L; MEIER, Paul. *Nonparametric estimation from incomplete observations*. **Journal of the American statistical association**, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

KLEIN, John P.; MOESCHBERGER, Melvin L. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd. New York: Springer, 2003. ISBN 978-0-387-95399-1.

LAWLESS, Jerald F. *Statistical models and methods for lifetime data*. New York: John Wiley & Sons, 1982.

MCCULLAGH, Peter; NELDER, J. A. *Generalized Linear Models*. London: Chapman e Hall, 1989.

NELSON, Wayne B. *Accelerated Life Testing: Statistical Models, data analysis and test plans*. New York: John Wiley & Sons, 1990.

PAGANO, M.; GAUVREAU, K. *Princípios de Bioestatística*. São Paulo: Cengage Learning, 2003.

PRENTICE, Ross L *et al. The analysis of failure times in the presence of competing risks*. **Biometrics**, JSTOR, p. 541–554, 1978.

R CORE TEAM. *R: A language and environment for statistical computing*. Vienna, 2024.

RAI, Rahul; OUTROS. *Survival analysis: A primer for clinician scientists*. **Journal of Postgraduate Medicine**, v. 68, n. 1, p. 1–8, 2022. DOI: 10.4103/jpgm.jpgm_553_21. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC8743691/.

WEIBULL, Waloddi. *A statistical theory of strength of materials*. **IVB-Handl.**, 1939.

# A. Anexos

## A.1 Algoritmos Utilizado na construção dos Resultados da Primeira Análise

**Listing A.1:** Simulação dos dados - Primeira Análise

```
library(survival)
set.seed(250203)
n<-100
dias_por_mes <- 30

# Simulacao Exponencial, Weibull e Log normal
tempos1<- rexp(n, rate = 0.05) # Exponencial
tempos2<- rweibull(n, shape = 2, scale = 10) # Weibull
tempos3<- rlnorm(n, meanlog = 2, sdlog = 0.5) # Log normal

# Simulacao Censura 50%
cens1<-sample(c(0,1),n, prob=c(0.5,0.5),replace=T)
cens2<-sample(c(0,1),n, prob=c(0.5,0.5),replace=T)
cens3<-sample(c(0,1),n, prob=c(0.5,0.5),replace=T)

dados<-data.frame(tempos1,cens1,tempos2,cens2, tempos3, cens3)
head(dados)
View(dados)
```

**Listing A.2:** Ajustes das simulações com Máxima Verossimilhança

```
# Ajuste Exponecial: maxima verossimilhanca
ekm1 <- survfit(Surv(tempos1 / dias_por_mes, cens1) ~ 1)
time1 <- ekm1$time
st1   <- ekm1$surv

ajust1 <- survreg(Surv(tempos1 / dias_por_mes, cens1) ~ 1, dist="
    exponential")
ajust1
alpha1 <- exp(ajust1$coefficients[1])
```

```r
alpha1

ste1   <- exp(-time1/alpha1)
res1 <- data.frame(time=time1, KM=st1, Exponencial=ste1)
head(res1)
View(res1)

# Ajuste weibull: maxima verossimilhanca
ekm2 <- survfit(Surv(tempos2 / dias_por_mes, cens2) ~ 1)
time2 <- ekm2$time
st2   <- ekm2$surv

ajust2 <- survreg(Surv(tempos2 / dias_por_mes, cens2) ~ 1, dist="weibull")
ajust2
alpha2 <- exp(ajust2$coefficients[1])  # escala
gama2  <- 1/ajust2$scale               # forma

stw2   <- exp(-(time2/alpha2)^gama2)
res2 <- data.frame(time=time2, KM=st2, Weibull=stw2)
head(res2)
View(res2)

# Ajuste Log normal: maxima verossimilhanca
ekm3 <- survfit(Surv(tempos3 / dias_por_mes, cens3) ~ 1)
time3 <- ekm3$time
st3   <- ekm3$surv

ajust3 <- survreg(Surv(tempos3 / dias_por_mes, cens3) ~ 1, dist="lognormal"
   )
ajust3
mu3    <- ajust3$coefficients[1]
sigma3 <- ajust3$scale

stln3  <- 1 - pnorm((log(time3) - mu3)/sigma3)
res3 <- data.frame(time=time3, KM=st3, LogNormal=stln3)
head(res3)
View(res3)
```

**Listing A.3:** Comparação dos parâmetros simulados e os estimados por máxima verossimilhança.

```r
library(dplyr)
# Parametros verdadeiros da simulacao
# Exponencial
rate1 <- 0.05
alpha1_true <- (1 / rate1) / dias_por_mes  # media verdadeira em meses

# Weibull
shape2_true <- 2
scale2_true <- 10 / dias_por_mes           # parametro de escala verdadeiro
    (em meses)

# Log-normal
meanlog3_true <- 2
sdlog3_true   <- 0.5
mu3_true      <- meanlog3_true - log(dias_por_mes)  # conversao para log(
   meses)
sigma3_true   <- sdlog3_true

# Parametros estimados dos ajustes ja feitos
# Exponencial
alpha1_est <- alpha1  # ja calculado anteriormente
se_alpha1  <- alpha1_est * summary(ajust1)$table[1, "Std. Error"]

# Weibull
alpha2_est <- alpha2
gama2_est  <- gama2

# Log-normal
mu3_est    <- mu3
sigma3_est <- sigma3
se_mu3     <- summary(ajust3)$table[1, "Std. Error"]

#Tabela comparando parametros verdadeiros e estimados

comparacao_parametros <- data.frame(
Modelo = c("Exponencial", "Weibull", "Weibull", "Log-normal", "Log-normal")
   ,
Parametro = c("alpha (escala)", "alpha (escala)", "gamma (forma)", "mu", "
   sigma"),
```

```r
Verdadeiro = c(alpha1_true, scale2_true, shape2_true, mu3_true, sigma3_true
    ),
Estimado   = c(alpha1_est, alpha2_est, gama2_est, mu3_est, sigma3_est)
)

# Calculo dos erros
comparacao_parametros <- comparacao_parametros %>%
mutate(
Erro_Absoluto = abs(Estimado - Verdadeiro),
Erro_Relativo_percent = round(100 * Erro_Absoluto / abs(Verdadeiro), 2)
)

# Visualizacao da tabela
print(comparacao_parametros)
```

**Listing A.4:** Gráficos ponto a ponto e Log Verossimilhança

```r
# Grafico comparacao ponto a ponto

par(mfrow=c(1,3))
# Exponencial
plot(st1, ste1, pch=16, ylim=c(0,1), xlim=c(0,1),
xlab = "S(t): Kaplan-Meier", ylab="S(t): Exponencial")
lines(c(0,1), c(0,1), lty=2)

# Weibull
plot(st2, stw2, pch=16, ylim=c(0,1), xlim=c(0,1),
xlab = "S(t): Kaplan-Meier", ylab="S(t): Weibull")
lines(c(0,1), c(0,1), lty=2)

# Log-Normal
plot(st3, stln3, pch=16, ylim=c(0,1), xlim=c(0,1),
xlab = "S(t): Kaplan-Meier", ylab="S(t): Log-Normal")
lines(c(0,1), c(0,1), lty=2)

# Log Verossimilhanca (quantitativa)
ajust1$loglik[2]
ajust2$loglik[2]
ajust3$loglik[2]
```

**Listing A.5:** Gráficos para avaliação dos modelos

```
# Grafico avaliacao dos modelos
par(mfrow=c(1,3))

plot(ekm1, conf.int = F, xlab = "Tempo (em meses)", ylab = "S(t)",
main = "Ajuste do Modelo Exponencial")
lines(c(0,time1), c(1,ste1), lty=2, col="purple") # Cor da curva
    Exponencial
legend("bottomleft", lty=c(1,2), col=c("black", "purple"),
c("Kaplan-Meier", "Expinencial"), bty="n", cex=0.8)

plot(ekm2, conf.int=F, xlab="Tempo (em meses)", ylab="S(t)",
main="Ajuste do Modelo Weibull")
lines(c(0,time2), c(1,stw2), lty=2, col="red") # Cor da curva Weibull
legend("bottomleft", lty=c(1,2), col=c("black", "red"),
c("Kaplan-Meier", "Weibull"), bty="n", cex=0.8)

plot(ekm3, conf.int=F, xlab="Tempo (em meses)", ylab="S(t)",
main="Ajuste do Modelo Log-normal")
lines(c(0,time3), c(1,stln3), lty=2, col="blue") # Cor da curva Log-normal
legend("bottomleft", lty=c(1,2), col=c("black", "blue"),
c("Kaplan-Meier", "Log-normal"), bty="n", cex=0.8)
```

**Listing A.6:** Gráfico dos Resíduos e seu algoritmo.

```
# Residuos de Cox-Snell
r_coxsnell_exp <- -log(1 - pexp(tempos1 / dias_por_mes, rate = 1/alpha1))
r_coxsnell_weib <- -log(1 - pweibull(tempos2 / dias_por_mes, shape = gama2,
     scale = alpha2))
r_coxsnell_ln <- -log(1 - plnorm(tempos3 / dias_por_mes, meanlog = mu3,
    sdlog = sigma3))

residuos <- data.frame(individuo = 1:n, exp = r_coxsnell_exp, weibull = r_
    coxsnell_weib,
lognormal = r_coxsnell_ln, cens1, cens2, cens3)

head(residuos)

# Exponencial
```

```r
ei <- r_coxsnell_exp
ekm_exp <- survfit(Surv(ei, cens1) ~ 1)
t_exp <- ekm_exp$time
st_exp <- ekm_exp$surv
sexp_exp <- exp(-t_exp)

par(mfrow=c(1,2))

# Grafico ponto a ponto
plot(st_exp, sexp_exp, xlab="S(res): Kaplan-Meier", ylab="S(res):
    Exponencial padrao",
pch=16, main=paste("Residuos de Cox-Snell: Exponencial"))
mtext("(a)", side=1, line=4, adj=0.5)

# Grafico avaliacao modelo
plot(ekm_exp, conf.int=F, mark.time=F, xlab="Residuos de Cox-Snell",
ylab="Sobrevivencia estimada")
lines(t_exp, sexp_exp, lty=4, col="purple")
legend("topright", lty=c(1,4), col=c("black","purple"),
legend=c("Kaplan-Meier","Exponencial padrao"), bty="n", cex=0.8)
mtext("(b)", side=1, line=4, adj=0.5)


# Weibull
ei <- r_coxsnell_weib
ekm_weib <- survfit(Surv(ei, cens2) ~ 1)
t_weib <- ekm_weib$time
st_weib <- ekm_weib$surv
sexp_weib <- exp(-t_weib)

par(mfrow=c(1,2))

# Grafico ponto a ponto
plot(st_weib, sexp_weib, xlab="S(res): Kaplan-Meier", ylab="S(res): Weibull
    padrao",
pch=16, main=paste("Residuos de Cox-Snell: Weibull"))
mtext("(a)", side=1, line=4, adj=0.5)

# Grafico avaliacao modelo
```

```
plot(ekm_weib, conf.int=F, mark.time=F, xlab="Residuos de Cox-Snell",
ylab="Sobrevivencia estimada")
lines(t_weib, sexp_weib, lty=4, col="red")
legend("topright", lty=c(1,4), col=c("black","red"),
legend=c("Kaplan-Meier","Exponencial padrao"), bty="n", cex=0.8)
mtext("(b)", side=1, line=4, adj=0.5)

# Log-normal
ei <- r_coxsnell_ln
ekm_ln <- survfit(Surv(ei, cens3) ~ 1)
t_ln <- ekm_ln$time
st_ln <- ekm_ln$surv
sexp_ln <- exp(-t_ln)

par(mfrow=c(1,2))

# Grafico ponto a ponto
plot(st_ln, sexp_ln, xlab="S(res): Kaplan-Meier", ylab="S(res): log-normal
    padrao",
pch=16, main=paste("Residuos de Cox-Snell: Log-normal"))
mtext("(a)", side=1, line=4, adj=0.5)

# Grafico avaliacao modelo
plot(ekm_ln, conf.int=F, mark.time=F, xlab="Residuos de Cox-Snell",
ylab="Sobrevivencia estimada")
lines(t_ln, sexp_ln, lty=4, col="blue")
legend("topright", lty=c(1,4), col=c("black","blue"),
legend=c("Kaplan-Meier","Exponencial padrao"), bty="n", cex=0.8)
mtext("(b)", side=1, line=4, adj=0.5)
```

## A.2 Algoritmos Utilizado na construção dos Resultados da Segunda Análise

**Listing A.7:** Simulação dos dados - Segunda Análise

```
library(survival)

```

```r
set.seed(250225)
n.pacientes <- 100
sexo <- sample(c("Homem", "Mulher"), size = n.pacientes, replace = TRUE)
tempo_max <- 12
idade <- round(rnorm(n.pacientes, mean = 60, sd = 12))
idade[idade < 30] <- 30 # minimo 30 anos
idade[idade > 90] <- 90 # maximo 90 anos

tempo_base <- ifelse(sexo == "Homem",
rweibull(sum(sexo == "Homem"), shape = 1.5, scale = 8),
rweibull(sum(sexo == "Mulher"), shape = 1.3, scale = 5))

tempo_ajustado <- tempo_base - (idade - 60) * 0.1
tempo_ajustado[tempo_ajustado < 0.1] <- runif(sum(tempo_ajustado < 0.1),
    0.1, 0.5)

tempo_final <- round(pmin(tempo_ajustado, tempo_max),1)

# Simulacao censura - 1 = evento (remissao) e 0 = censura (morte/evasao)
status <- ifelse(runif(n.pacientes) < 0.8 & tempo_final < tempo_max, 1, 0)

dados <- data.frame( id = 1:n.pacientes, sexo = sexo, idade = idade, tempo
    = tempo_final,
status = status)
View(dados)
```

**Listing A.8:** Plot gráfico Kaplan-Meier e Teste Logrank.

```r
# Curva de Kaplan-Meier
ekm <- survfit(Surv(tempo_final, status) ~ sexo, data = dados)
summary(ekm)

teste_logrank <- survdiff(Surv(tempo_final, status) ~ sexo, rho = 0)
teste_logrank

par(pin = c(6, 4))
par(mar = c(5, 5, 4, 2) + 0.1)
plot(ekm, lty = c(2, 1), main = "Estimador de Kaplan-Meier", cex.main =
    1.0,
xlab = "Tempo (meses)", ylab = "S(t) estimada",
```

```r
col = c("blue", "red"), mark.time = TRUE)

text(x = 10, y = 0.8, expression(p == 0.03), cex = 1)
legend("bottomleft", lty = c(2, 1), legend = c("Homem", "Mulher"),
lwd = 1, bty = "n", col = c("blue", "red"))
```

**Listing A.9:** Modelos de Regressão Paramétrico.

```r
# Modelos de Regressao Parametricos
# Modelo Exponencial
mod_exp <- survreg(Surv(tempo, status) ~ sexo + idade, data = dados, dist =
    "exponential")

# Modelo Weibull
mod_weib <- survreg(Surv(tempo, status) ~ sexo + idade, data = dados, dist
   = "weibull")

# Modelo Log-Normal
mod_lnorm <- survreg(Surv(tempo, status) ~ sexo + idade, data = dados, dist
    = "lognormal")

# Comparacao dos modelos
AIC(mod_exp, mod_weib, mod_lnorm)

summary(mod_exp)
summary(mod_weib)
summary(mod_lnorm)
```

**Listing A.10:** Gráfico dos Modelos Regressão Paramétrico.

```r
# Sequencia de tempo
tempo_seq <- seq(0, tempo_max, by = 0.1)

# Criacao dos data frames para cada sexo (idade media = 60)
novo_homem  <- data.frame(sexo = "Homem",  idade = 60)
novo_mulher <- data.frame(sexo = "Mulher", idade = 60)

# Modelos Parametricos por sexo
# Exponencial
S_exp_homem  <- 1 - pexp(tempo_seq, rate = 1 / exp(predict(mod_exp, newdata
    = novo_homem,  type = "lp")))
```

```r
S_exp_mulher <- 1 - pexp(tempo_seq, rate = 1 / exp(predict(mod_exp, newdata
    = novo_mulher, type = "lp")))

# Weibull
S_weib_homem  <- 1 - pweibull(tempo_seq, shape = 1 / mod_weib$scale,
scale = exp(predict(mod_weib, newdata = novo_homem,  type = "lp")))
S_weib_mulher <- 1 - pweibull(tempo_seq, shape = 1 / mod_weib$scale,
scale = exp(predict(mod_weib, newdata = novo_mulher, type = "lp")))

# Log-Normal
S_lnorm_homem  <- 1 - plnorm(tempo_seq, meanlog = predict(mod_lnorm,
   newdata = novo_homem,  type = "lp"),
sdlog = mod_lnorm$scale)
S_lnorm_mulher <- 1 - plnorm(tempo_seq, meanlog = predict(mod_lnorm,
   newdata = novo_mulher, type = "lp"),
sdlog = mod_lnorm$scale)

# Graficos lado a lado
par(mfrow = c(1,3), mar = c(5,4,4,2))

# Exponencial
plot(ekm, lty = c(2,1), col = c("blue","red"), mark.time = TRUE,
main = "Exponencial", xlab = "Tempo", ylab = "S(t)")
lines(tempo_seq, S_exp_homem,  col = "blue",  lwd = 2)
lines(tempo_seq, S_exp_mulher, col = "red",   lwd = 2)
legend("bottomleft", legend = c("KM Homem", "KM Mulher", "Exp Homem", "Exp
   Mulher"),
col = c("blue","red","blue","red"), lty = c(2,1,1,1), lwd = 2, bty = "n")

# Weibull
plot(ekm, lty = c(2,1), col = c("blue","red"), mark.time = TRUE,
main = "Weibull", xlab = "Tempo", ylab = "S(t)")
lines(tempo_seq, S_weib_homem,  col = "blue",  lwd = 2)
lines(tempo_seq, S_weib_mulher, col = "red",   lwd = 2)
legend("bottomleft", legend = c("KM Homem", "KM Mulher", "Weibull Homem", "
   Weibull Mulher"),
col = c("blue","red","blue","red"), lty = c(2,1,1,1), lwd = 2, bty = "n")

# Log-Normal
```

```r
plot(ekm, lty = c(2,1), col = c("blue","red"), mark.time = TRUE,
main = "Log-Normal", xlab = "Tempo", ylab = "S(t)")
lines(tempo_seq, S_lnorm_homem,  col = "blue",  lwd = 2)
lines(tempo_seq, S_lnorm_mulher, col = "red",   lwd = 2)
legend("bottomleft", legend = c("KM Homem", "KM Mulher", "Log-Normal Homem"
    , "Log-Normal Mulher"),
col = c("blue","red","blue","red"), lty = c(2,1,1,1), lwd = 2, bty = "n")
```

**Listing A.11:** Coeficientes dos modelos aplicados aos dados simualdas.

```r
# Extraindo coeficientes dos modelos
coef_exp  <- coef(mod_exp)
coef_weib <- coef(mod_weib)
coef_lnorm <- coef(mod_lnorm)

# Calculando intercepto efetivo para cada grupo
# Exponencial
intercepto_exp_h  <- coef_exp["(Intercept)"]
intercepto_exp_m  <- coef_exp["(Intercept)"] + coef_exp["sexoMulher"]
beta_idade_exp    <- coef_exp["idade"]

# Weibull
intercepto_weib_h <- coef_weib["(Intercept)"]
intercepto_weib_m <- coef_weib["(Intercept)"] + coef_weib["sexoMulher"]
beta_idade_weib   <- coef_weib["idade"]

# Log-Normal
intercepto_lnorm_h <- coef_lnorm["(Intercept)"]
intercepto_lnorm_m <- coef_lnorm["(Intercept)"] + coef_lnorm["sexoMulher"]
beta_idade_lnorm   <- coef_lnorm["idade"]

# Tabela final
tabela_betas <- data.frame(
Modelo = rep(c("Exponencial", "Weibull", "Log-Normal"), each = 2),
Grupo = rep(c("Homens", "Mulheres"), 3),
Intercepto = c(intercepto_exp_h, intercepto_exp_m,
intercepto_weib_h, intercepto_weib_m,
intercepto_lnorm_h, intercepto_lnorm_m),
Beta_idade = c(beta_idade_exp, beta_idade_exp,
beta_idade_weib, beta_idade_weib,
```

```
31 beta_idade_lnorm, beta_idade_lnorm))
32
33 tabela_betas
```