Análise de Sobrevivência

Iniciação Ciêntifica - PIBIC 2024/2025 (UFPA)

Breno Cauã Rodrigues da Silva
Prof Dr Paulo Cerqueira dos Santos Jr

Índice

Prefácio

Este é um projeto desenvolvido...

1 Conceitos Básicos e Exemplos

1.1 Introdução

O objetivo deste capítulo inicial é apresentar alguns *conceitos* e *fundamentos* de uma das áreas da Estatística e Análise de Dados que mais se desenvolveram nas últimas duas décadas do século XX. Esse avanço foi impulsionado pela evolução das técnicas estatísticas aliada ao progresso computacional.

Na Análise de Sobrevivência, a variável resposta é, em geral, o *tempo até a ocorrência de um evento de interesse*. Especificamente, essa área se concentra em modelar e compreender o tempo necessário para que um evento significativo ocorra, sendo este denominado **tempo de falha**. Como exemplo, Colosimo e Giolo (2006) mencionam casos como o tempo até a morte de um paciente, até a cura de uma doença ou até a recidiva de uma condição clínica.

Uma questão frequentemente levantada é: por que não utilizar outras técnicas estatísticas? Métodos tradicionais não são adequados para dados de sobrevivência devido a uma característica única: a **censura**. Esse conceito refere-se à observação parcial do tempo de falha, como ocorre quando o acompanhamento de um paciente é interrompido antes do evento de interesse. A censura, sendo um elemento essencial da Análise de Sobrevivência, caracteriza situações em que o tempo de falha real é desconhecido, sabendo-se apenas que ele excede determinado ponto.

1.2 Tempo de Falha

Na Análise de Sobrevivência, é fundamental estabelecer alguns pontos iniciais para o estudo. O primeiro deles é o **tempo inicial do estudo**, que deve ser claramente definido para garantir que os indivíduos sejam comparáveis no ponto de partida, diferenciando-se apenas pelas covariáveis medidas. Existem diversas maneiras de definir o tempo inicial, sendo o mais comum o **tempo cronológico**. Contudo, em áreas como Engenharia, outras métricas, como número de ciclos ou quilometragem, também podem ser utilizadas. Colosimo e Giolo (2006) apresentam exemplos práticos, como medidas de carga para equipamentos.

Outro aspecto essencial é a **definição do evento de interesse**, frequentemente associado a falhas ou situações indesejáveis. Para garantir resultados consistentes, a definição do evento deve ser clara e objetiva. Um exemplo elucidativo é fornecido por Colosimo e Giolo (2006):

> *"Em algumas situações, a definição de falha já é clara, como morte ou recidiva, mas em outras pode assumir termos ambíguos. Por exemplo, fabricantes de produtos alimentícios desejam saber o tempo de vida de seus produtos expostos em balcões frigoríficos de supermercados. O tempo de falha vai do momento de exposição (chegada ao supermercado) até o produto se tornar 'inapropriado para*

consumo'. Esse evento deve ser claramente definido antes do início do estudo. Por exemplo, o produto é considerado inapropriado para consumo quando atinge uma concentração específica de microrganismos por mm^2 de área."

1.3 Censura

Estudos clínicos que tratam a resposta como uma variável temporal geralmente são prospectivos e de longa duração. No entanto, mesmo sendo extensos, esses estudos frequentemente se encerram antes que todos os indivíduos passem pelo evento de interesse.

Uma característica comum nesses estudos é a **censura**, que corresponde a observações incompletas ou parciais. Apesar disso, tais observações fornecem informações valiosas para a análise. Colosimo e Giolo (2006) destacam a relevância de incluir dados censurados na análise:

> *"Ressalta-se que, mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser incluídos na análise estatística. Duas razões justificam esse procedimento: (i) mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de vida dos pacientes; (ii) a exclusão das censuras no cálculo das estatísticas pode levar a conclusões enviesadas."*

Existem três tipos principais de censura:

- **Censura Tipo I:** O estudo é encerrado após um período de tempo previamente definido.
- **Censura Tipo II:** O estudo termina quando um número específico de indivíduos passa pelo evento de interesse.
- **Censura Aleatória:** Ocorre quando um indivíduo é retirado do estudo antes do evento de interesse.

A censura mais comum é a **censura à direita**, em que o evento ocorre após o tempo registrado. Entretanto, outros tipos de censura, como **à esquerda** e **intervalar**, também são possíveis.

Censura à esquerda ocorre quando o evento já aconteceu antes do início da observação. Um exemplo é um estudo sobre a idade em que crianças aprendem a ler:

> *"Quando os pesquisadores começaram a pesquisa, algumas crianças já sabiam ler e não se lembravam com que idade isso ocorreu, caracterizando observações censuradas à esquerda."*

No mesmo estudo, observa-se censura à direita para crianças que ainda não sabiam ler no momento da coleta de dados. Nesse caso, os tempos de vida são classificados como **duplamente censurados** (Turnbull 1974).

A censura intervalar ocorre em estudos com visitas periódicas espaçadas, onde só se sabe que o evento ocorreu dentro de um intervalo de tempo. Quando o tempo de falha T é impreciso, considera-se que ele pertence a um intervalo $T \in (L, U]$, conhecido como **sobrevivência**

intervalar. Casos especiais incluem tempos de falha exatos, em que $L = U$, sendo $U = 0$ para censura à direita e $L = 0$ para censura à esquerda (Lindsey e Ryan 1998). Destaca-se a seguinte observação de Colosimo e Giolo (2006):

> *"A presença de censura traz desafios para a análise estatística. A censura do Tipo II é, em princípio, mais tratável que os outros tipos, mas para situações simples, que raramente ocorrem em estudos clínicos (Lawless 1982). Na prática, utiliza-se resultados assintóticos para a análise dos dados de sobrevivência."*

1.4 Dados Truncados

O truncamento é uma característica de alguns estudos de sobrevivência que, muitas vezes, é confundida com a censura. Ele ocorre quando certos indivíduos são excluídos do estudo devido a uma condição específica. Nesse caso, os pacientes só são incluídos no acompanhamento após passarem por um determinado evento, em vez de serem acompanhados desde o início do processo.

1.5 Representação dos Dados de Sobrevivência

Considere uma amostra aleatória de tamanho n. O i-ésimo indivíduo no estudo é geralmente representado pelo par (t_i, δ_i), onde t_i é o tempo de falha ou censura, indicado pela variável binária δ_i, definida como:

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

Portanto, a variável resposta na análise de sobrevivência é representada por duas colunas no conjunto de dados. Se o estudo também incluir covariáveis, os dados são representados por $(t_i, \delta_i, \mathbf{x}_i)$. Caso a censura seja intervalar, a representação é $(li, u_i, \delta_i, \mathbf{x}_i)$. Para exemplos de dados de sobrevivência, veja a Seção 1.5 do livro de Colosimo e Giolo (2006).

1.6 Especificando o Tempo de Sobrevivência

Seja T uma variável aleatória (v.a.), na maioria dos casos contínua, que representa o tempo de falha. Assim, o suporte de T é definido nos reais positivos $\mathbb{R}^+$. Tal variável é geralmente representada pela sua *função risco* ou pela *função de taxa de falha* (ou taxa de risco). Tais funções, e outras relacionadas, são usadas ao longo do processo de análise de dados de sobrevivência. A seguir, algumas dessas funções e as relações entre elas serão definidas.

1.6.1 Função de Sobrevivência

Esta é uma das principais funções probabilísticas usadas em análise de sobrevivência. A função sobrevivência é definida como a probabilidade de uma observação não falhar até certo

ponto t, ou seja a probabilidade de uma observação sobreviver ao tempo t. Em probabilidade, isso pode ser escrito como:

$$S(t) = P(T > t), \tag{1.1}$$

uma conclusão a qual podemos chegar, é que a probabilidade de uma observação não sobreviver até o tempo t, é a acumulada até o ponto t, logo,

$$F(t) = 1 - S(t). \tag{1.2}$$

1.6.2 Função de Taxa de Falha ou de Risco

A probabilidade da falha ocorrer em um intervalo de tempo $[t_1, t_2)$ pode ser expressa em termos da função de sobrevivência como:

$$S(t_1) - S(t_2).$$

A taxa de falha no intervalo $[t_1, t_2)$ é definida como a probabilidade de que a falha ocorra neste intervalo, dado que não ocorreu antes de t_1, dividida pelo comprimento do intervalo. Assim, a taxa de falha no intervalo $[t_1, t_2)$ é expressa por

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}.$$

De forma geral, redefinindo o intervalo como $[t, t + \Delta t)$ a expressão assume a seguinte forma:

$$\lambda(t) = \frac{S(t) - S(t + \Delta_t)}{\Delta t \ S(t)}.$$

Assumindo Δt bem pequeno, $\lambda(t)$ representa a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t. Observe que as taxas de falha são números positivos, mas sem limite superior. A função de taxa de falha $\lambda(t)$ é bastante útil para descrever a distribuição do tempo de vida de pacientes. Ela descreve a forma em que a taxa instantânea de falha muda com o tempo. A função de taxa de falha de T é, então, definida como:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}. \tag{1.3}$$

A função de taxa de falha é mais informativa do que a função de sobrevivência. Diferentes funções de sobrevivência podem ter formas semelhantes, enquanto as respectivas funções de taxa de falha podem diferir drasticamente. Desta forma, a modelagem da função de taxa de falha é um importante método para dados de sobrevivência.

1.6.3 Função de Taxa de Falha Acumulada

Outra função útil em análise de dados de sobrevivência é a função taxa de falha acumulada. Esta função, como o próprio nome sugere, fornece a taxa de falha acumulada do indivíduo e é definida por:

$$\Lambda(t) = \int_0^t \lambda(u)du. \tag{1.4}$$

A função de taxa de falha acumulada, $\Lambda(t)$, não têm uma interpretação direta, mas pode ser útil na avaliação da função de maior interesse que é a função de taxa de falha, $\lambda(t)$. Isto acontece essencialmente na estimação não-paramétrica em que $\Lambda(t)$ apresenta um estimador com propriedades ótimas e $\lambda(t)$ é difícil de ser estimada.

1.6.4 Tempo Médio e Vida Média Residual

Outras duas quantidades de interesse em análise de sobrevivência são: o tempo médio de via e a vida média residual. A primeira é obtida pela área sob a função de sobrevivência. Isto é,

$$t_m = \int_0^\infty S(t)dt. \tag{1.5}$$

Já a vida média residual é definida condicional a um certo tempo de vida t. Ou seja, para indivíduos com idade t está quantidade mede o tempo médio restante de vida e é, então, a área sob a curva de sobrevivência à direita do tempo t dividida por $S(t)$. Isto é,

$$\text{vmr}(t) = \frac{\int_0^\infty (u-t)f(u)du}{S(t)} = \frac{\int_0^\infty S(u)du}{S(t)}, \tag{1.6}$$

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

1.7 Relações entre as Funções

Para T uma variável aleatória contínua e não-negativa, tem-se, em termos das funções definidas anteriormente, algumas relações matemáticas importantes entre elas, a saber:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\left[\log S(t)\right],$$

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

e

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

Tais relações mostram que o conhecimento de uma das funções, por exemplo $S(t)$, implica no conhecimento das demais, isto é, $F(t)$, $f(t)$, $\lambda(t)$ e $\Lambda(t)$. Outras relações envolvendo estas funções são as seguintes:

$$S(t) = \frac{\text{vmr}(0)}{\text{vmr}(t)} \exp\left\{ -\int_0^t \frac{du}{\text{vmr}(u)} \right\}$$

e

$$\lambda(t) = \left(\frac{d\ [\text{vmr}(t)]}{dt} + 1 \right) / \text{vmr}(t).$$

2 Técnicas Não Paramétricas

2.1 Introdução

Este capítulo apresenta as técnicas não-paramétricas utilizadas para a análise de dados de sobrevivência. Essas técnicas são empregadas quando não se faz suposições sobre a forma específica da distribuição dos tempos de falha, sendo particularmente úteis para dados censurados.

2.2 O Estimador de Kaplan-Meier

Proposto por Kaplan e Meier (1958). É um estimador não-paramétrico utilizado para estimar a função de sobrevivência, $S(t)$. Tal estimador também é chamado de de *estimador limite-produto*. O Estimador de Kaplan-Meier é uma adaptação a $S(t)$ empiríca que, na ausência de censura nos dados, é definida como:

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

$\hat{S}(t)$ é uma função que tem um formato gráfico de escada com degraus nos tempos observados de falha de tamanho $1/n$, onde n é o tamanho amostral.

O processo utilizado até se obter a estimativa de Kaplan-Meier é um processo passo a passo, em que o próximo passo depende do anterior. De forma suscetível, para qualquer t, $S(t)$ pode ser escrito em termos de probabilidades condicionais. Suponha que existam n pacientes no estudo e $k(\leq n)$ falhas distintas nos tempos $t_1 \leq t_2 \leq \cdots \leq t_k$. Considerando $S(t)$ uma função discreta com probabilidade maior que zero somente nos tempos de falha t_j, $j = 1, \cdots, k$, tem-se que:

$$S(t_j) = (1 - q_1)(1 - q_2)\cdots(1 - q_j), \tag{2.1}$$

em que q_j é a probabilidade de um indivíduo morrer no intervalo $[t_{j-1}, tj)$ sabendo que ele não morreu até t_{j-1} e considerando $t_0 = 0$. Ou seja, pode se escrever q_j como:

$$q_j = P(T \in [t_{j-1}, t_j) | T \geq t_{j-1}), \tag{2.2}$$

para $j = 1, \cdots, k$.

A expressão geral do estimador de Kaplan-Meier pode ser apresentada após estas considerações preliminares, Formalmente, considere:

- $t_1 \leq t_2 \leq \cdots \leq t_k$, os k tempos distintos e ordenados de falha;
- d_j o número de falhas em t_j, $j = 1, \cdots, k$;

- n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

Com isso, pode-se definir o estimador de Kaplan-Meier como:

$$\hat{S}_{KM}(t) = \prod_{j\,:\,t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j\,:\,t_j<t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.3}$$

De forma intuitiva, por assim dizer, a Equação 2.3 é proveniente da Equação 2.1, sendo está, uma decomposição de $S(t)$ em termos q_j's. Assim, a Equação 2.3 é justificada se os q_j's forem estimados por d_j/n_j, que em palavras está expresso na Equação 2.2. No artigo original de 1958, Kaplan e Meier provam que a Equação 2.3 é um *Estimador de Máxima Verossimilhança* (EMV) para $S(t)$. Seguindo certos passos, é possível provar que que $\hat{S}_{KM}(t)$ é EMV de $S(t)$. Supondo que d_j observações falham no tempo tempo t_j, para $j = 1, \cdots, k$, e m_j observações são censuradas no intervalo $[t_j, t_{j+1})$, nos tempos $t_{j1}, \ldots, t_{jm_j}$. A probabilidade de falha no tempo t_j é, então,

$$S(t_j) - S(t_j+),$$

com $S(t_j+) = \lim_{\Delta t \to 0+} S(t_j + \Delta t)$, $j = 1, \cdots, k$. Por outro lado, a contribuição para a função de verossimilhança de um tempo de sobrevivência censurado em t_{jl} para $l = 1, \ldots, m_j$, é:

$$P(T > t_{jl}) = S(t_{jl}+).$$

A função de verossimilhança pode, então, ser escrita como:

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

Com isso, é possível provar que $S(t)$ que maximiza $L(S(\cdot))$ é exatamente a expressão dada pela Equação 2.3.

2.2.1 Propriedades do Estimador de Kaplan-Meier

Como um estimador de máxima verossimilhança, o estimador de Kaplan-Meier têm interessantes propriedades. As principais são:

- É não-viciado para grandes amostras;
- É fracamente consistente;
- Converge assintoticamente para um processo gaussiano.

A consistência e normalidade assintótica de $\hat{S}_{KM}(t)$ foram provadas sob certas condições de regularidade, por Breslow e Crowley (1974) e Meier (1975) e, no artigo original, Kaplan e Meier (1958) mostram que $\hat{S}_{KM}(t)$ é um EMV para $S(t)$, como já dito.

2.2.2 Variância do Estimador de Kaplan-Meier

Para que se possa construir intervalos de confiança e testar hipóteses para $S(t)$, se faz necessário ter conhecimento quanto variabilidade e precisão do estimador de Kaplan-Meier. Este estimador, assim como outros, está sujeito a variações que devem ser descritas em termos de estimações intervalares. A expressão da variância assintótica do estimador de Kaplan-Meier é dada pela Equação 2.4.

$$\hat{Var}[\hat{S}_{KM}(t)] = [\hat{S}_{KM}(t)]^2 \sum_{j\,:\,t_j<t} \frac{d_j}{n_j(n_j - d_j)} \tag{2.4}$$

A expressão dada na Equação 2.4, é conhecida como fórmula de Greenwood e pode ser obtida a partir de propriedades do estimador de máxima verossimilhança. Os detalhes da obtenção da Equação 2.4 estão disponíveis em Kalbfleisch e Prentice (1980).

Como $\hat{S}_{KM}(t)$, para um t fixo, tem distribuição assintóticamente Normal. O intervalo de confiança com $100(1-\alpha)\%$ de confiança para $\hat{S}_{KM}(t)$ é expresso por:

$$\hat{S}_{KM}(t) \pm z_{\alpha/2}\sqrt{\hat{Var}[\hat{S}_{KM}(t)]}.$$

Vale salientar que para valores extremos de t, este intervalo de confiança pode apresentar limites que não condizem com a teoria de probabilidades. Para solucionar tal problema, aplica-se uma transformação em $S(t)$ como, por exemplo, $\hat{U}(t) = \log\left[-\log\left(\hat{S}_{KM}(t)\right)\right]$. Esta transformação foi sugerida por Kalbfleisch e Prentice (1980), tendo sua variância estimada por:

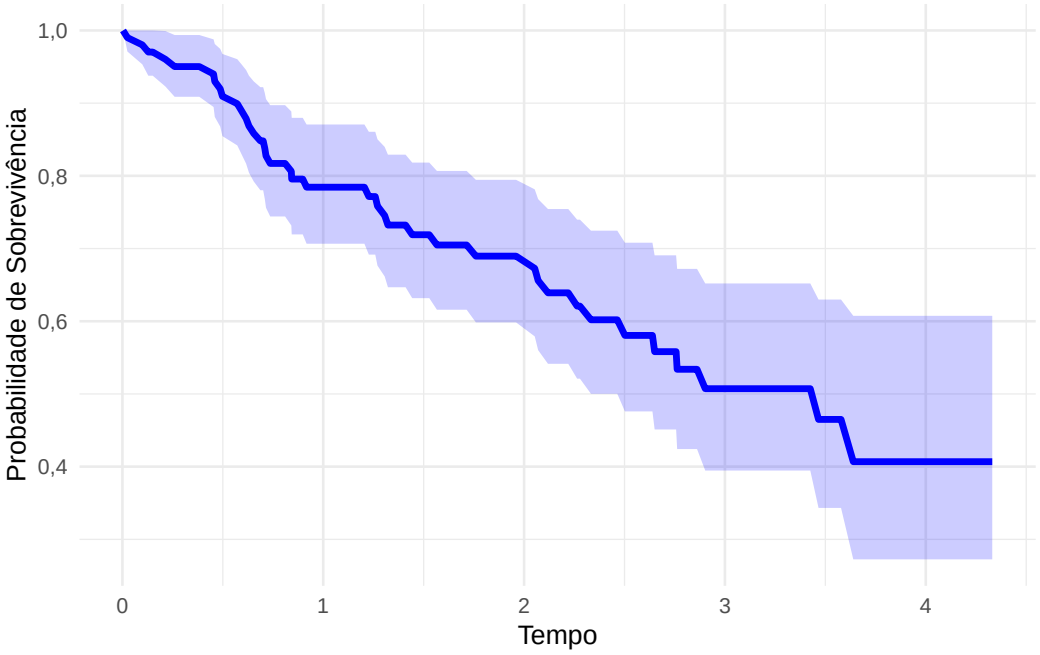
$$\hat{Var}[\hat{U}(t)] = \frac{\sum_{j\,:\,t_j<t} \dfrac{d_j}{n_j(n_j-d_j)}}{\left[\sum_{j\,:\,t_j<t} \log\left(\dfrac{n_j - d_j}{n_j}\right)\right]^2} = \frac{\sum_{j\,:\,t_j<t} \dfrac{d_j}{n_j(n_j-d_j)}}{\left[\log \hat{S}_{KM}(t)\right]^2}$$

Logo, pode-se aproximar um intervalo com $100(1-\alpha)\%$ de confiança para $S(t)$ desta forma:

$$\left[\hat{S}(t)\right]^{\exp\left\{\pm z_{\alpha/2}\sqrt{\hat{Var}[\hat{U}(t)]}\right\}}.$$

Veja uma aplicação do estimador de Kaplan-Meier para os dados de *Leucemia Pediátrica* dispostos no Apêndice (A) do livro *Análise de Sobrevivência Aplicada* de Colosimo e Giolo (2006). De posse do conjunto de dados, pode-se estimar a curva de sobrevivência, tal curva foi ilustrada na Figura 2.1.

Figura 2.1: Curva de Sobrevivência de Kaplan-Meier com IC de 95%



2.3 Outros Estimadores Não Parâmetricos

O estimador de Kaplan-Meier é, indiscutivelmente, o mais utilizado para estimar $S(t)$ em análises de sobrevivência. Ele é amplamente disponibilizado em diversos pacotes estatísticos e abordado em inúmeros textos de estatística básica. Entretanto, outros dois estimadores de $S(t)$ também possuem relevância significativa na literatura especializada: o estimador de Nelson-Aalen e o estimador da tabela de vida.

O estimador de Nelson-Aalen, mais recente que o de Kaplan-Meier, apresenta propriedades similares às deste último. Já o estimador da tabela de vida possui importância histórica, tendo sido utilizado em informações derivadas de censos demográficos para estimar características associadas ao tempo de vida humano. Este estimador foi inicialmente proposto por demógrafos e atuários no final do século XIX, sendo empregado principalmente em grandes amostras.

Nesta seção será abordado apenas o estimador de Nelson-Aalen. Para conhecer mais sobre o estimador da Tabela de Vida ou Tabela Atuarial, consulte a Seção 2.4.2 do livro *Análise de Sobrevivência Aplicada* de Colosimo e Giolo (2006).

2.3.1 Estimador de Nelson-Aalen

Mais recente que o estimador de Kaplan-Meier, este estimador se baseia na função de sobrevivência expressa da seguinte forma:

$$S(t) = \exp\{-\Lambda(t)\},$$

em que $\Lambda(t)$ é a função de risco acumulado apresentada na Seção 1.6.3.

A estimativa para $\Lambda(t)$ foi inicialmente proposta por Nelson (1972) posteriormente retomada por Aalen (1978) que demonstrou suas propriedades assintóticas utilizando processos de contagem. Na literatura, esse estimador é amplamente conhecido como o estimador de Nelson-Aalen e é definido pela seguinte expressão:

$$\hat{\Lambda}(t) = \sum_{j:t_j<t} \left(\frac{d_j}{n_j}\right), \tag{2.5}$$

onde d_j e n_j são as mesmas definições usadas no estimador de Kaplan-Meier. A variância do estimador, conforme proposta por Aalen (1978), é dada por:

$$\hat{Var}(\hat{\Lambda}(t)) = \sum_{j:t_j<t} \left(\frac{d_j}{n_j^2}\right). \tag{2.6}$$

Uma alternativa para a estimativa da variância de $\hat{\Lambda}(t)$, proposta por Klein (1991), é:

$$\hat{Var}(\hat{\Lambda}(t)) = \sum_{j:t_j<t} \frac{(n_j - d_j)d_j}{n_j^3},$$

entretanto, o estimador da Equação 2.6 apresenta menor vício, tornando-o mais preferível que o proposto por Klein (1991).

Desta forma, podemos definir, com base no estimador de Nelson-Aalen, um estimador para a função de sobrevivência, podendo ser expressa por:

$$\hat{S}_{NA}(t) = \exp\left\{-\hat{\Lambda}(t)\right\}.$$

Deve-se, a variância deste estimador, a Aalen e Johansen (1978). Podendo ser mensurada pela expressão:

$$\hat{Var}(\hat{S}_{NA}(t)) = \left[\hat{S}_{NA}(t)\right]^2 \sum_{j:t_j<t} \left(\frac{d_j}{n_j^2}\right)$$

Vale destacar que o estimador de Nelson-Aalen apresenta, na maioria dos casos, estimativas próximas ao estimador de Kaplan-Meier. Bohoris (1994) mostrou que $\hat{S}_{NA}(t) \geq \hat{S}_{KM}(t)$ para todo t, isto é, as estimativas obtidas pelo estimador de Nelson-Aalen são maiores ou iguais às estimativas obtidas pelo estimador de Kaplan-Meier.

2.4 Comparação de Curvas de Sobrevivência

Considere um problema na área da saúde em que se deseja comparar dois grupos: um que receberá tratamento com uma determinada droga e outro que será o grupo controle. Estatísticas amplamente utilizadas para esse fim podem ser vistas como generalizações, para dados censurados, de testes não paramétricos bem conhecidos. Entre esses, o teste *logrank* (Mantel 1966) é o mais empregado em análises de sobrevivência. Gehan (1965) propôs uma generalização para a estatística de Wilcoxon. Outras generalizações foram introduzidas por

autores como Peto e Peto (1972) e Prentice (1978), enquanto Latta (1981) utilizou simulações de Monte Carlo para comparar diversos testes não-paramétricos.

Nesta seção, será dada ênfase ao teste *logrank*, amplamente utilizado em análises de sobrevivência e particularmente adequado quando a razão entre as funções de risco dos grupos a serem comparados é aproximadamente constante. Ou seja, quando as populações apresentam a propriedade de riscos proporcionais.

A estatística do teste *logrank* baseia-se na diferença entre o número observado de falhas em cada grupo e o número esperado de falhas sob a hipótese nula. Essa abordagem é semelhante à do teste de Mantel e Haenszel (1959), que combina tabelas de contingência. Além disso, o teste *logrank* possui a mesma expressão do teste de escore para o modelo de regressão de Cox.

Considere, inicialmente, o teste de igualdade entre duas funções de sobrevivência $S_1(t)$ e $S_2(t)$. Seja $t_1 < t_2 < ... < t_k$ a sequência dos tempos de falha distintos observados na amostra combinada, formada pela união das duas amostras individuais. Suponha que, no tempo t_j, ocorram d_j falhas e que n_j indivíduos estejam sob risco imediatamente antes de t_j na amostra combinada. Nas amostras individuais, as quantidades correspondentes são d_{ij} e n_{ij}, onde $i = 1, 2$ representa o grupo e $j = 1, \cdots, k$ indica o tempo de falha.

No tempo t_j, os dados podem ser organizados em uma tabela de contingência 2×2, onde d_{ij} representa o número de falhas e $n_{ij} - d_{ij}$ o número de sobreviventes em cada grupo i. Essa disposição está ilustrada na Tabela 2.1.

Tabela 2.1: Tabela de contingência gerada no tempo t_j.

	Grupo 1	Grupo 2	
Falha	d_{1j}	d_{2j}	d_j
Não Falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
	n_{1j}	n_{2j}	n_j

Condicionado à ocorrência de falhas e censuras até o tempo t_j (fixando as marginais das colunas) e ao número total de falhas no tempo t_j (fixando as marginais das linhas), a distribuição de d_{2j} é, então, uma hipergeométrica:

$$\frac{\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}}{\binom{n_j}{d_j}}.$$

A média de d_{2j} é dada por $w_{2j} = n_{2j}d_j n_j^{-1}$. Isso significa que, na ausência de diferenças entre as duas populações no tempo t_j, o número total de falhas (d_j) pode ser alocado entre as duas amostras proporcionalmente à razão entre o número de indivíduos sob risco em cada amostra e o número total sob risco.

A variância de d_{2j} obtida a partir da distribuição hipergeométrica é:

$$(V_j)_2 = n_{2j}(n_j - n_{2j})d_j(n_j - d_j)n_j^{-2}(n_j - 1)^{-1}.$$

Portanto, a estatística $d_{2j} - w_{2j}$ possui média zero e variância $(V_j)_2$. Se as k tabelas de contingência forem independentes, um teste aproximado para avaliar a igualdade entre as duas funções de sobrevivência pode ser construído com base na seguinte estatística:

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2}, \tag{2.7}$$

que, sob a hipótese nula $H_0 : S_1(t) = S_2(t)$ para todo t no período de acompanhamento, segue aproximadamente uma distribuição qui-quadrado com 1 grau de liberdade para amostras grandes.

Para exemplificar a aplicação do teste de *logrank* em dados reais, utilizou-se o conjunto de dados sobre Leucemia Pediátrica, disponível no Apêndice (A) do livro *Análise de Sobrevivência Aplicada* de Colosimo e Giolo (2006). Esses mesmos dados foram usados para gerar a Figura 2.1. O objetivo do teste realizado foi avaliar se as curvas de sobrevivência das categorias da covariável `r6` são iguais, com as seguintes hipóteses:

$$\begin{cases} H_0 : \text{As curvas de sobrevivência dos grupos são iguais ao longo do tempo} \\ H_1 : \text{As curvas de sobrevivência dos grupos são diferentes ao longo do tempo.} \end{cases}$$

Veja a saída resultante do teste realizado no software R:

```
Call:
survdiff(formula = Surv(tempos, cens) ~ grupo, data = dados,
    rho = 0)

                     N Observed Expected (O-E)^2/E (O-E)^2/V
grupo=Category One  95       34    37,16     0,269      5,73
grupo=Category Zero  8        5     1,84     5,429      5,73

 Chisq= 5,7  on 1 degrees of freedom, p= 0,02
```

Ao fixar o nível de significância em 5% ($\alpha = 0,05$), rejeitamos a hipótese nula. Essa conclusão baseia-se no valor p (probabilidade de significância) obtido no teste, calculado como $p - valor = 0{,}0166$. Como o $p - valor < \alpha$, rejeita-se H_0. Assim, conclui-se que as curvas de sobrevivência dos grupos são diferentes ao longo do tempo, ao nível de significância de 5%.

A generalização do teste *logrank* para a comparação de $r > 2$ funções de sobrevivência, $S_1(t), S_2(t), \ldots, S_r(t)$, é direta. Utilizando a mesma notação anterior, o índice i varia agora de 1 a r. Assim, os dados podem ser organizados em uma tabela de contingência $2 \times r$, onde cada coluna i contém d_{ij} falhas e $n_{ij} - d_{ij}$ sobreviventes. Dessa forma, a Tabela 2.1 seria estendida para ter r colunas em vez de apenas duas.

Condicionada à experiência de falha e censura até o tempo t_j e ao número total de falhas no tempo t_j, a distribuição conjunta de $d_{2j}, \ldots, d_{rj}$ segue uma hipergeométrica multivariada, dada por:

$$\frac{\prod_{i=1}^{r}\binom{n_{ij}}{d_{ij}}}{\binom{n_j}{d_j}}.$$

A média de d_{ij} é $w_{ij} = n_{ij}d_j n_j^{-1}$, bem como a variância de d_{ij} e a covariância de d_{ij} e d_{lj} são, respectivamente,

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

e

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

A estatística $v_j' = (d_{2j} - w_{2j}, \ldots, d_{rj} - w_{rj})$ possui média zero e matriz de variância-covariância V_j, com dimensão $r - 1$. A matriz V_j contém os termos $(V_j)ii$ na diagonal principal e $(V_j)_{il}$, $i, l = 2, \ldots, r$, fora da diagonal principal.

A estatística v, que agrega as contribuições de todos os tempos distintos de falha, é definida como:

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

onde v é um vetor de dimensão $(r - 1) \times 1$, cujos elementos correspondem às diferenças entre os totais observados e esperados de falhas.

Considerando, novamente, a independência das k tabelas de contingência, a variância de v é dada por $V = V_1 + \ldots + V_k$. Um teste aproximado para a igualdade das r funções de sobrevivência pode ser baseado na estatística:

$$T = v´V^{-1}v, \tag{2.8}$$

que, sob a hipótese nula H_0 (igualdade das curvas de sobrevivência), segue uma distribuição qui-quadrado com $r - 1$ graus de liberdade para amostras grandes. Os graus de liberdade são $r - 1$ em vez de r, pois os elementos de v somam zero.

Uma aplicação para a comparação de r curvas de sobrevivência...

`Código em R a ser preenchido`

2.4.1 Outros Testes

[...]

3 Técnicas Paramétricas - Modelos Probabilísticos

3.1 Introdução

No capítulo anterior, foi apresentada uma abordagem não paramétrica para a análise de dados de sobrevivência, na qual a estimação é realizada sem assumir uma distribuição de probabilidade específica para o tempo de sobrevivência.

Os estimadores não paramétricos são derivados diretamente do conjunto de dados, pressupondo que o mecanismo gerador das informações opera de maneira distinta em diferentes momentos no tempo, funcionando de forma quase independente. Assim, conclui-se que a abordagem não paramétrica possui tantos parâmetros quanto intervalos de tempo considerados. Contudo, ao incluir covariáveis, o modelo de Kaplan-Meier não permite estimar diretamente o "efeito" dessas covariáveis, limitando-se a comparar e testar a igualdade entre diferentes curvas de sobrevivência.

Por outro lado, nos modelos de regressão tradicionais, como os modelos *linear*, *Poisson* ou *logístico*, a escolha de uma distribuição de probabilidade para a variável resposta Y e de uma função para a relação entre Y e as covariáveis $x_1, x_2, \ldots, x_p$ é essencial para identificar o modelo. Ao aplicar esse conceito na análise de sobrevivência, o tempo até a ocorrência de um evento de interesse é tratado como a variável resposta.

Nesse contexto, este capítulo introduz uma abordagem paramétrica para estimar as funções básicas de sobrevivência. Assume-se que a distribuição de probabilidade do tempo de ocorrência do evento é conhecida, permitindo a estimação dos parâmetros associados ao modelo de forma mais estruturada e eficiente.

3.2 Distribuições do Tempo de Sobrevivência

Seja T uma variável aleatória que representa o "tempo de sobrevivência". Qual seria a distribuição de probabilidade mais adequada para representá-la?

Uma característica fundamental da variável aleatória T é que ela é contínua e não negativa. Com base nessa propriedade, é possível eliminar algumas distribuições como candidatas adequadas para modelar T. Por exemplo, a distribuição normal não é apropriada, pois admite valores negativos, o que contradiz a natureza do tempo de sobrevivência. Além disso, os tempos de sobrevivência frequentemente apresentam uma forte assimetria à direita, reforçando a inadequação da distribuição normal para esse contexto.

3.2.1 Distribuição Exponencial

Se $T \sim Exp(\alpha)$, a sua função densidade de probabilidade é expressa da seguinte forma:

$$f(t) = \alpha \exp\{-\alpha t\},\ t \geq 0 \text{ e } \alpha > 0. \tag{3.1}$$

Desta forma, podemos obter a função de sobrevivência com base no completar da distribuição acumulada de T:

$$\begin{aligned} S(t) &= P(T > t) = 1 - P(T \leq t) = 1 - F(t) \\ &= 1 - [1 - \exp\{-\alpha t\}] \\ &= \exp\{-\alpha t\}. \end{aligned}$$

Assim definimos, formalmente, a função de sobrevivência como:

$$S(t) = \exp\{-\alpha t\}. \tag{3.2}$$

Note que o parâmetro α é a velocidade de queda da função sobrevivência. Através das relações entre as funções em análise de sobrevivência, temos a função risco ou taxa de falha. Obtida pela razão entre da função densidade de probabilidade e a função de sobrevivência:

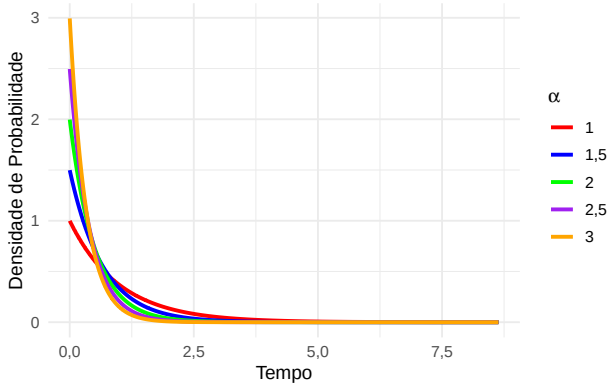
$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\alpha \exp\{-\alpha t\}}{\exp\{-\alpha t\}} = \alpha = \text{constante}. \tag{3.3}$$

Sendo a função risco constante para todo tempo observado t, o risco acumulado é função linear no tempo com uma inclinação da reta dada por α:

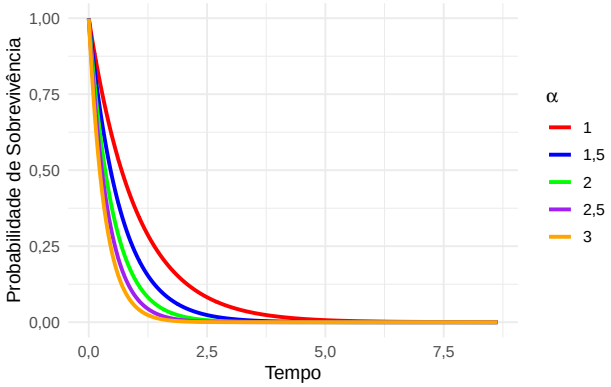
$$\Lambda(t) = -\ln[S(t)] = -\ln[\exp\{-\alpha t\}] = -(-\alpha t) = \alpha t \tag{3.4}$$

Veja, a seguir, a Figura 3.1 que mostra as curvas de densidade de probabilidade, de sobrevivência, risco e risco acumulado para diferentes valores do parâmetro α.

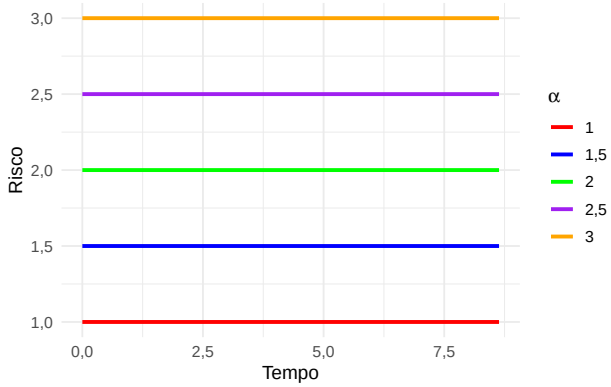
Figura 3.1: Funções Densidade de Probabilidade, Sobrevivência, Risco e Risco Acumulado segundo uma Distribuição Exponencial para diferentes valores do Parâmetro de Taxa.



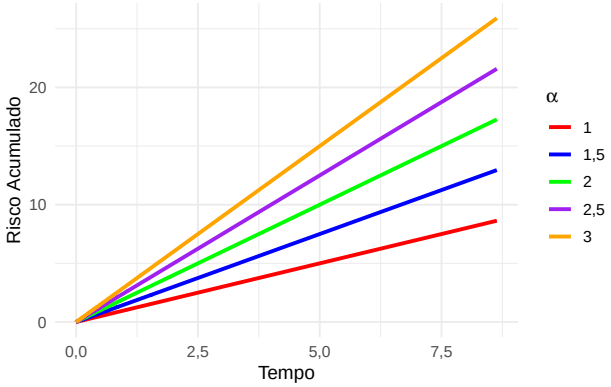
(a) Função Densidade de Probabilidade



(b) Função de Sobrevivência



(c) Função de Risco



(d) Função de Risco Acumulado

3.2.1.1 Algumas considerações

Note que, quanto maior o valor de α (risco), mais abruptamente a função de sobrevivência $S(t)$ decresce, e maior é a inclinação da função de risco acumulado.

A distribuição exponencial, por possuir um único parâmetro, é matematicamente simples e apresenta um formato assimétrico. Seu uso em análise de sobrevivência tem uma analogia com a suposição de normalidade em outras técnicas e áreas da estatística. Entretanto, a suposição de risco constante associada a essa distribuição é bastante restritiva e, em muitos casos, pode não ser realista.

Por exemplo, considere um estudo sobre câncer, em que o tempo até o evento de interesse é definido como o período até a morte ou a cura do paciente. Para aplicar a distribuição exponencial nesse contexto, seria necessário assumir que o tempo desde o diagnóstico da doença não afeta a probabilidade de ocorrência do evento. Essa suposição é delicada, pois o próprio passar do tempo afeta naturalmente a probabilidade de sobrevivência, o risco e o risco acumulado, entre outros fatores. Isso pode ocorrer por causas naturais, como o envelhecimento, que aumenta o risco com o avanço da idade. Essa característica da distribuição exponencial é conhecida como falta de memória, o que significa que o risco futuro é independente do tempo

já decorrido.

Quando $\alpha = 1$, a distribuição é denominada exponencial padrão. A média e a variância do tempo de sobrevivência, para uma variável que segue a distribuição exponencial, são expressas como funções inversas do parâmetro de risco (α). Assim, quanto maior o risco, menor o tempo médio de sobrevivência e menor a variabilidade em torno da média. As expressões são dadas por:

$$E[T] = \frac{1}{\alpha},$$

$$Var[T] = \frac{1}{\alpha^2}.$$

Como a distribuição de T é assimétrica, se torna mais usual utilizar o *tempo mediano de sobrevivência* ao invés de tempo médio. Pode-se obter o tempo mediano de sobrevivência a partir de um tempo t, tal que, $S(t) = 0,5$, logo,

$$\begin{aligned} S(t) = 0,5 \Leftrightarrow \exp\{-\alpha t\} = 0,5 \Leftrightarrow -\alpha t = \ln(2^{-1}) \\ \alpha t = -[-\ln(2)] \Leftrightarrow \alpha t = \ln(2). \end{aligned}$$

Desta forma, o tempo mediano de sobrevivência é definido como:

$$T_{mediano} = \frac{\ln(2)}{\alpha}.$$

Em resumo, o modelo exponencial é apropriado para situações em que o período do experimento é curto o suficiente para que a suposição de risco constante seja plausível.

3.2.2 Distribuição Weibull

Na maioria dos casos de análise de sobrevivência na área da saúde, é mais razoável supor que o risco varia ao longo do tempo, em vez de permanecer constante.

Atualmente, a *Distribuição Weibull* é amplamente utilizada, pois permite modelar essa variação do risco ao longo do tempo. Como será demonstrado, a distribuição exponencial é um caso particular da distribuição Weibull.

Se o tempo de sobrevivência T segue uma distribuição Weibull, ou seja, $T \sim Weibull(\gamma, \alpha)$, sua função densidade de probabilidade é dada por:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}. \tag{3.5}$$

A partir da Equação 3.5 é possível chegar a função de sobrevivência da distribuição Weibull sendo está função definida como:

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}, \tag{3.6}$$

onde $t \geq 0$, α o parâmetro escala (ou taxa) e γ parâmetro de forma. Ambos os parâmetros sempre positivos.

A função de risco, $\lambda(t)$, depende do tempo de sobrevivência. Apresentando variação no tempo conforme a expressão:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \tag{3.7}$$

e a função de risco acumulado da distribuição Weibull é dada por:

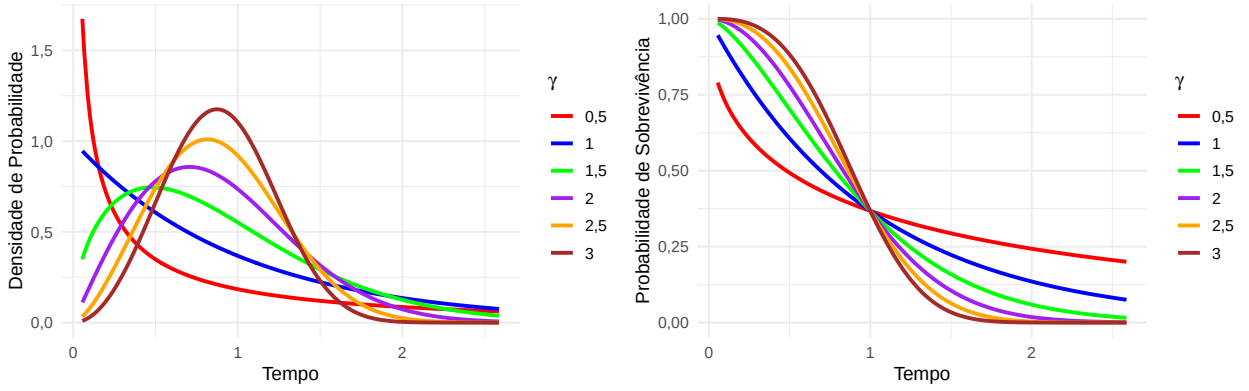
$$\Lambda(t) = -\ln[S(t)] = -\ln\left[\exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}\right] = \left(\frac{t}{\alpha}\right)^{\gamma}. \tag{3.8}$$

Note que, o parâmetro γ determina a forma função de risco da seguinte maneira:

- $\gamma < 1 \rightarrow$ função de risco decresce;
- $\gamma > 1 \rightarrow$ função de risco cresce;
- $\gamma = 1 \rightarrow$ função de risco constante, caindo no caso particular da distribuição exponencial.

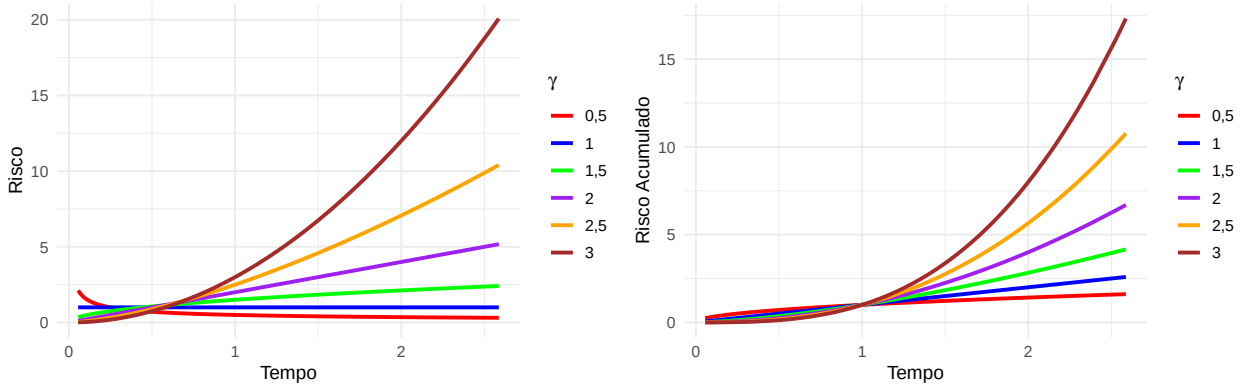
Veja, a seguir, a Figura 3.2 que mostra as curvas de densidade, sobrevivência, risco e risco acumulado para diferentes valores do parâmetro de forma γ e o de escala $\alpha = 1$.

Figura 3.2: Funções Densidade de Probabilidade, Sobrevivência, Risco e Risco Acumulado segundo uma Distribuição Weibull para diferentes valores do parâmetro de forma.



(a) Função Densidade de Probabilidade

(b) Função de Sobrevivência



(c) Função de Risco

(d) Função de Risco Acumulado

3.2.2.1 Algumas considerações

É incluso a função gama na média e variância da distribuição Weibull, assim,

$$E[T] = \alpha\Gamma[1 + (1/\gamma)]$$

e

$$Var[T] = \alpha^2 \left[\Gamma[1 + (2/\gamma)] - \Gamma[1 + (1/\gamma)]^2\right]$$

sendo a função gama $\Gamma[k]$, expressa por $\Gamma[k] = \int_0^\infty t^{k-1} \exp\{t\}dt$.

Afim de se obter o tempo mediano de sobrevivência, igualamos a probabilidade de sobrevivência a $0,5$. Desta forma:

$$\begin{aligned} S(t) = 0,5 \Leftrightarrow \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\} &= 0,5 \\ -\left(\frac{t}{\alpha}\right)^\gamma = \ln\left(2^{-1}\right) \Leftrightarrow \left(\frac{t}{\alpha}\right)^\gamma &= \ln(2) \\ \frac{t}{\alpha} &= [\ln(2)]^{1/\gamma}. \end{aligned}$$

Logo, definimos o tempo mediano de sobrevivência da distribuição Weibull como:

$$T_{mediano} = \alpha[\ln(2)]^{1/\gamma}.$$

3.2.3 Distribuição Log-normal

Uma outra possibilidade para modelar o tempo de sobrevivência é a *distribuição Log-normal*. Dizer que $T \sim Normal(\mu, \sigma^2)$ implica em dizer que $\ln(T) \sim Log - normal(\mu, \sigma^2)$ em que μ é a média do logaritmo do tempo de falha e σ^2 sua variância. Pode-se fazer uso desta relação para modelar o tempo de sobrevivência conforme uma distribuição normal, desde que, se aplique o logaritmo aos dados observados. A função densidade para tal distribuição é dada por:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln(t) - \mu}{\sigma}\right)^2\right\}. \tag{3.9}$$

Assim, quando o tempo de sobrevivência segue uma distribuição log-normal, sua função de sobrevivência e as demais não tem uma forma análitica explícita, desde modo, deve-se fazer uso das relações entre as funções para se obter a função taxa de falha e taxa de falha acumulada. Desta forma, essas funções são expressas, respectivamente, por:

$$S(t) = \Phi\left(\frac{-\ln(t) + \mu}{\sigma}\right), \tag{3.10}$$

$$\lambda(t) = \frac{f(t)}{S(t)}$$

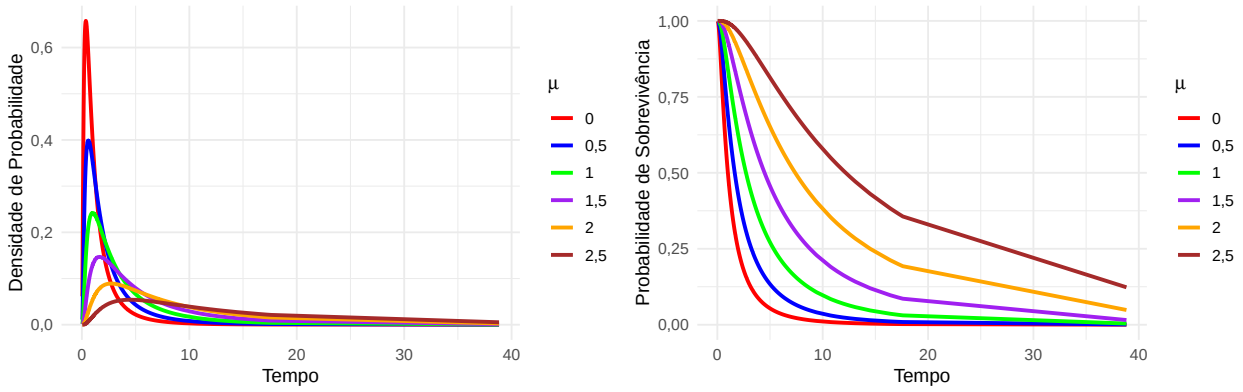
e

$$\Lambda(t) = -\ln[S(t)]$$

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

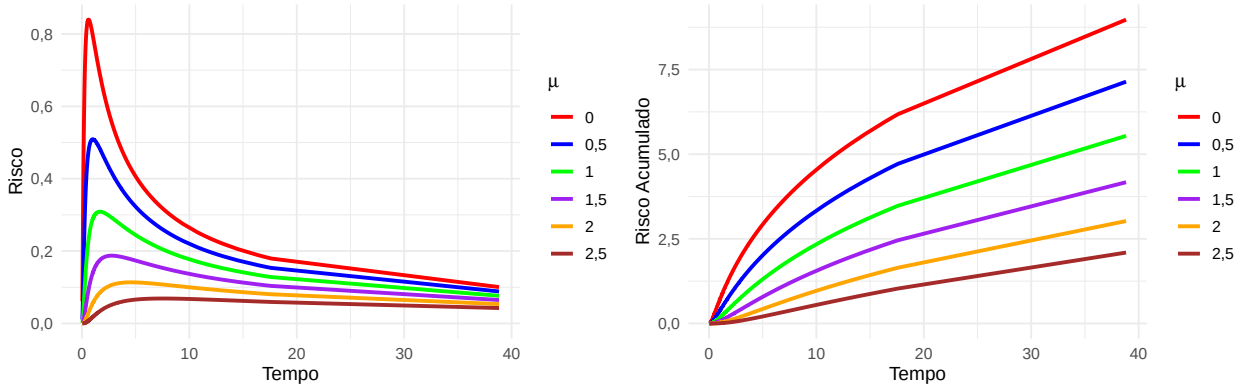
Veja a Figura 3.3 que ilustras as curvas usadas na análise de sobrevivência segundo uma distribuição log-normal, variando o parâmetro de locação μ e fixando o parâmetro de escala $\sigma = 1$.

Figura 3.3: Funções Densidade de Probabilidade, Sobrevivência, Risco e Risco Acumulado segundo uma Distribuição Log-normal para diferentes valores do parâmetro de média.



(a) Função Densidade de Probabilidade

(b) Função de Sobrevivência



(c) Função de Risco

(d) Função de Risco Acumulado

3.2.3.1 Algumas considerações

A média e a variância da distribuição log-normal são, respectivamente, dadas por:

$$E[T] = \exp\{\mu + \sigma^2/2\}$$

e

$$Var[T] = \exp\{2\mu + \sigma^2\}(\exp\{\sigma^2 - 1\})$$

3.2.4 Distribuição Exponencial por Partes

3.2.5 Distribuição Exponencial por Partes de Potência

3.3 Estimação de Parâmetros

Foram apresentados alguns modelos probabilísticos. Esses modelos possuem quantidades desconhecidas, denominadas **parâmetros**, ou **parâmetro**, quando o modelo depende de uma única quantidade desconhecida, como no caso da distribuição exponencial.

3.3.1 Método de Máxima Verossimilhança

O *Método de Máxima Verossimilhança* baseia-se no princípio de que, a partir de uma amostra aleatória, a melhor estimativa para o parâmetro de interesse é aquela que maximiza a probabilidade daquela amostra observada ter sido observada (Bussab e Morettin 2010).

De forma simples, o método de máxima verossimilhança condensa toda a informação contida na amostra, por meio da **função de verossimilhança**, para encontrar o(s) parâmetro(s) da distribuição que melhor expliquem os dados. Essa abordagem utiliza o produtório das densidades $f(t)$ para cada observação t_i, $i = 1, 2, \ldots, n$. Em livros introdutórios de estatística, a função de verossimilhança é definida da seguinte maneira, para um parâmetro ou vetor de parâmetros θ:

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

Observe que L é uma função de θ, que pode ser um único parâmetro ou um vetor de parâmetros, como ocorre na distribuição log-normal, onde $\theta = (\mu, \sigma^2)$. No entanto, em análise de sobrevivência, essa definição tradicional de função de verossimilhança é insuficiente, pois os dados frequentemente apresentam **censura**, o que implica que o tempo de evento pode ser apenas parcialmente observado.

Para lidar com essa característica, utiliza-se a variável indicadora δ_i, apresentada na Seção 1.5, que identifica se o i-ésimo tempo é um tempo de evento ou de censura. Com base nessa informação, a função de verossimilhança é ajustada da seguinte forma:

- Para $\delta_i = 1$, o i-ésimo tempo é um tempo de evento, e sua contribuição para $L(\theta)$ é a densidade de probabilidade $f(t_i|\theta)$;
- Para $\delta_i = 0$, o i-ésimo tempo é um tempo censurado, e sua contribuição para $L(\theta)$ é a função de sobrevivência $S(t_i|\theta)$.

Assim, a função de verossimilhança ajustada, que incorpora dados censurados, é expressa como:

$$L(\theta) = \prod_{i=1}^{n} \left[f(t_i|\theta)\right]^{\delta_i} \left[S(t_i|\theta)\right]^{1-\delta_i}$$

$$L(\theta) = \prod_{i=1}^{n} \left[\lambda(t_i|\theta)\right]^{\delta_i} S(t_i|\theta). \tag{3.11}$$

Para encontrar o valor de θ que maximiza $L(\theta)$, utiliza-se a derivada do logaritmo de base neperiana da verossimilhança igualada a zero:

$$\frac{\partial \ln[L(\theta)]}{\partial \theta} = 0.$$

A solução dessa equação fornece o valor de θ que maximiza $\ln[L(\theta)]$, e consequentemente, $L(\theta)$.

3.3.2 Método Iterativo de Newton-Raphson

Para algumas distribuições, apresentadas na Seção 3.2, e outras denifidas na literatura, não há forma analítica para as estimativas de máxima verossimilhança. Assim, as estimativas de tais parâmetros depende de métodos numéricos, sendo o **Método Iterativo de Newton-Raphson** uma abordagem amplamente utilizada.

O Método de Newton-Raphson é um procedimento iterativo eficiente para resolver equações não lineares, muito empregado na estimação de parâmetros de modelos estatísticos. No ajuste de distribuições o método busca maximizar a função de verossimilhança resolvendo o sistema de equações derivado das condições de otimalidade (gradiente nulo). A fórmula iterativa é:

$$\theta_{n+1} = \theta_n - \mathbf{H}^{-1}(\theta_n)\nabla \ln[L(\theta_n)], \tag{3.12}$$

onde:

- θ_n é o vetor de parâmetros estimados na iteração n;
- $\ln[L(\theta_n)]$ é o vetor gradiente, contendo as derivadas parciais de $\ln[L(\theta_n)]$ em relação as coordenadas do vetor θ (parâmetros);
- $\mathbf{H}(\theta)$ é a matriz Hessiana, composta pelas segundas derivadas de $\ln[L(\theta_n)]$.

O método apresenta vantagens convenientes no ajuste de parâmetros de modelos estatísticos. Uma das vantagens é a *eficiência* do método, que apresenta convergência rápida quando o ponto inicial θ_0 está próximo dos valores reais dos parâmetros. Outra vantagem, é *flexibilidade*, pois pode ser aplicado a diversos modelos probabilísticos, como o modelo Weibull, que é amplamente utilizada para modelar tempos de vida e dados de sobrevivência.

Entretanto, deve-se, também, atentar-se aos cuidados na aplicação do método. Pois, a *convergência* do método não é garantida caso o ponto inicial esteja muito distante da solução ou se as condições de regularidade do modelo não forem atendidas. Outro ponto que merece atenção é o cálculo da *matriz Hessiana*, que pode ser computacionalmente custoso, especialmente em modelos com maior complexidade.

Para um melhor entendimento do Método Iterativo de Newton-Raphson veja o Apêndice (D) do livro *Análise de Sobrevivência Aplicada* de Colosimo e Giolo (2006).

3.3.3 Aplicações Caso Não Haja Censura

Nesta seção, será demonstrado como determinar o estimador ou os estimadores de máxima verossimilhança para os parâmetros das distribuições discutidas quando não há presença de censuras nos dados. Aqui, será apresentada apenas a saída dos programas. Para ter o script utilizado, acesse o repositório do Github por meio do seguinte link: github.com

3.3.3.1 Distribuição Exponencial

Considere a distribuição exponencial conforme descrita na Seção 3.2.1. O **Estimador de Máxima Verossimilhança (EMV)** do parâmetro α, isto é, $\theta = \alpha$, pode ser obtido seguindo os passos descritos a seguir:

1. Definir a Função de Verossimilhança $L(\theta)$:

$$L(\theta) = \prod_{i=1}^{n} \left[\alpha \exp\{-\alpha t\}\right]^{\delta_i} \left[\exp\{-\alpha t\}\right]^{1-\delta_i}. \tag{3.13}$$

2. Tomar o logaritmo natural da função verossimilhança $\ln[L(\alpha)]$:

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

3. Derivar a função do log-verossimilhança em relação a θ. Logo $\dfrac{\partial \ln[L(\theta)]}{\partial \theta}$:

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

4. Igualar a derivada a zero e resolver para α:

$$\begin{aligned}\frac{\partial \ln[L(\theta)]}{\partial \theta} &= 0 \\ \frac{1}{\hat{\alpha}} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} t_i &= 0 \\ \hat{\alpha} &= \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}\end{aligned}$$

Note que, para o caso em que não se tem censura o numerador, $\sum_{i=1}^{n} \delta_i$, equivale ao tamanho da amostra n. Logo, o EMV para α no caso de não haver censura nos dados é: $\hat{\alpha} = n / \sum_{i=1}^{n} t_i$.

Simulou-se uma amostra proveniente de uma distribuição exponencial e, a partir dessa amostra, obteve-se a estimativa de máxima verossimilhança para o parâmetro α. Veja a

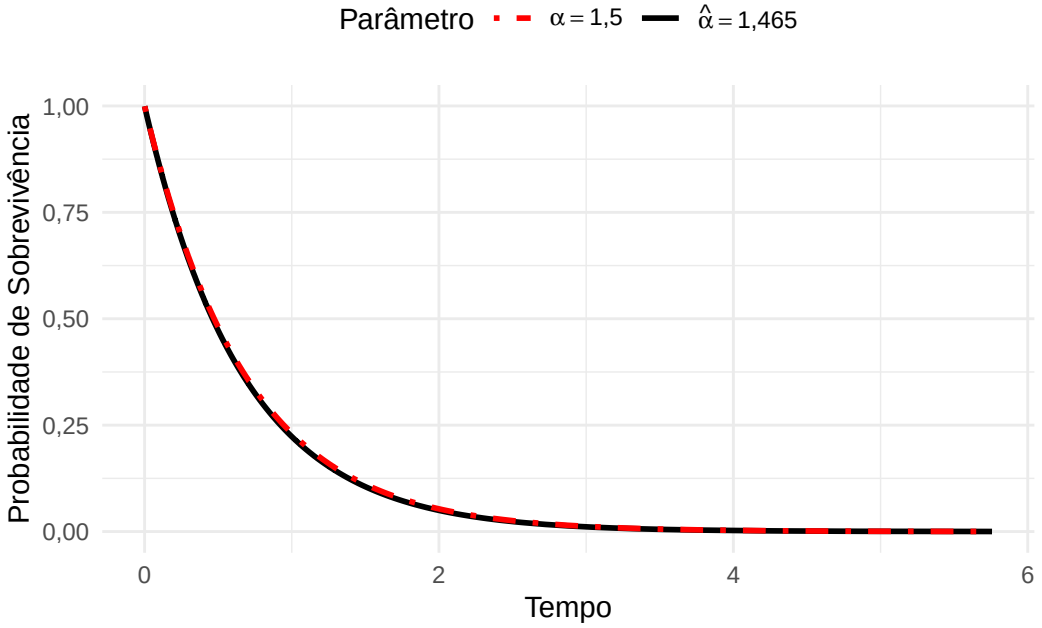
Tabela 3.1, que apresenta as dez observações, na ordem de simulação, e suas respectivas funções de sobrevivência real e estimada.

Tabela 3.1: Comparação de Dez Observações entre o valor Real e o Estimado por Máxima Verossimilhança da Função de Sobrevivência segundo o Modelo Exponencial.

Tempo	$S(t)$	$\hat{S}_{EMV}(t)$
0,2576	0,6795	0,6856
0,2305	0,7077	0,7134
0,2052	0,7351	0,7403
0,2920	0,6453	0,6519
0,1086	0,8497	0,8529
0,1239	0,8304	0,8339
2,1030	0,0427	0,0459
1,0528	0,2062	0,2138
0,4921	0,4780	0,4862
1,9384	0,0546	0,0584

O valor verdadeiro do parâmetro é $\alpha = 1{,}5$. A estimativa de máxima verossimilhança obtida foi $\hat{\alpha} = 1{,}4654$. Na Figura 3.4, comparamos graficamente as duas curvas de sobrevivência, ilustrando o valor real do parâmetro α e sua estimativa $\hat{\alpha}$.

Figura 3.4: Comparação das Curvas de Sobrevivência Real e Estimada por Máxima Verossimilhança segundo o Modelo Exponencial.



3.3.3.2 Distribuição Weibull

Para a estimação de parâmetros do modelo Weibull. Necessita-se da aplicação do método numérico de Newton-Raphson descrito na Seção 3.2. Tal método requer o cálculo das derivadas parciais e de segunda ordem em relação ao vetor de parâmetros $\theta = (\gamma, \alpha)$, permitindo ajustar o modelo aos dados observados de tempos de sobrevivência de forma precisa e eficiente.

Para o modelo Weibull será apresentada duas formas de implementar o método. A primeira é a construção braçal do algoritmo, que consiste na definição e cálculo explícito das funções necessárias, como a função de log-verossimilhança, o gradiente e a Hessiana. A segunda é utilizar a função de otimização `optim`, já implementadas na linguagem de programação **R**. Está função automatiza o processo de otimização e oferece uma implementação flexível e eficiente. Para os demais modelos probabilísticos será usado apenas a função de otimização `optim` do **R**.

Começando com a implementação sem uso da função de otimização, deve-se primeiramente, encontrar log-verossimilhança, o gradiente e a Hessiana. Pode-se definir a função de verossimilhança para distribuição Weibul usando a Equação 3.11, substituindo a função densidade e a função de sobrevivência da distribuição Weibull especificadas na Seção 3.2.2. Portanto:

$$L(\theta) = \prod_{i=1}^{n} \left[\frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\left\{ -\left(\frac{t}{\alpha}\right)^{\gamma} \right\} \right]^{\delta_i} \left[\exp\left\{ -\left(\frac{t}{\alpha}\right)^{\gamma} \right\} \right]^{1-\delta_i}. \tag{3.14}$$

Toma-se o logaritmo natural de $L(\gamma, \alpha)$, logo:

$$\begin{aligned} \ln[L(\gamma,\alpha)] &= \sum_{i=1}^{n} \delta_i \ln[\gamma] - \sum_{i=1}^{n} \delta_i \gamma \ln[\alpha] + \sum_{i=1}^{n} \delta_i (\gamma - 1) \ln[t_i] + \sum_{i=1}^{n} -(\alpha^{-1} t_i)^{\gamma} \\ &= \ln[\gamma] \sum_{i=1}^{n} \delta_i - \gamma \ln[\alpha] \sum_{i=1}^{n} \delta_i + (\gamma - 1) \sum_{i=1}^{n} \delta_i \ln[t_i] + \sum_{i=1}^{n} -(\alpha^{-1} t_i)^{\gamma}. \end{aligned}$$

Aplicando as derivadas de primeira ordem em relação a γ e α, temos:

$$\frac{\partial \ln[L(\gamma,\alpha)]}{\partial \gamma} = \frac{1}{\gamma} \sum_{i=1}^{n} \delta_i - \ln[\alpha] \sum_{i=1}^{n} \delta_i + \sum_{i=1}^{n} \delta_i \ln[t_i] - \sum_{i=1}^{n} (\alpha^{-1} t_i)^{\gamma} \ln[\alpha^{-1} t_i]$$

e

$$\frac{\partial \ln[L(\gamma,\alpha)]}{\partial \alpha} = -\frac{\gamma}{\alpha} \sum_{i=1}^{n} \delta_i + \gamma \alpha^{-\gamma-1} \sum_{i=1}^{n} t_i^{\gamma}.$$

Toma-se agora as derivadas de segunda ordem.

$$\frac{\partial^2 \ln[L(\gamma,\alpha)]}{\partial \gamma^2} = -\frac{1}{\gamma^2} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} (\alpha^{-1} t_i)^{\gamma} (\ln[\alpha^{-1} t_i])^2,$$

$$\frac{\partial^2 \ln[L(\gamma,\alpha)]}{\partial \alpha^2} = -\frac{\gamma}{\alpha^2} \sum_{i=1}^{n} \delta_i - \gamma(\gamma+1) \alpha^{-\gamma-2} \sum_{i=1}^{n} t_i^{\gamma}$$

e

$$\frac{\partial^2 \ln[L(\gamma,\alpha)]}{\partial\gamma\partial\alpha} = \frac{\partial^2 \ln[L(\gamma,\alpha)]}{\partial\alpha\partial\gamma} = -\frac{1}{\alpha}\sum_{i=1}^{n}\delta_i + \alpha^{-\gamma-1}\sum_{i=1}^{n} t_i^{\gamma}\left(\gamma \ln\left[\frac{t_i}{\alpha}\right] + 1\right)$$

Com todas as derivadas definidas, pode-se construir o algoritmo iterativo de Newton-Raphson. Tirou-se uma amostra de uma $Weibull(2;1,5)$. As estimativas obtidas foram as seguintes.

```
Número de Iterações Necessárias: 47

Estimativa para o parâmetro de forma: 1,964

Estimativa para o parâmetro de forma: 1,508
```

O mesmo resultado, ou bem próximo, pode ser obtido de uma forma mais direta por meio do uso da função `optim` para otimização. Veja a saída obtida de tal função.

```
$par
[1] 1,964 1,508

$value
[1] 1017

$counts
function gradient
      21        6

$convergence
[1] 0

$message
NULL

$hessian
       [,1]   [,2]
[1,]  477,5 -282,7
[2,] -282,7 1696,4
```

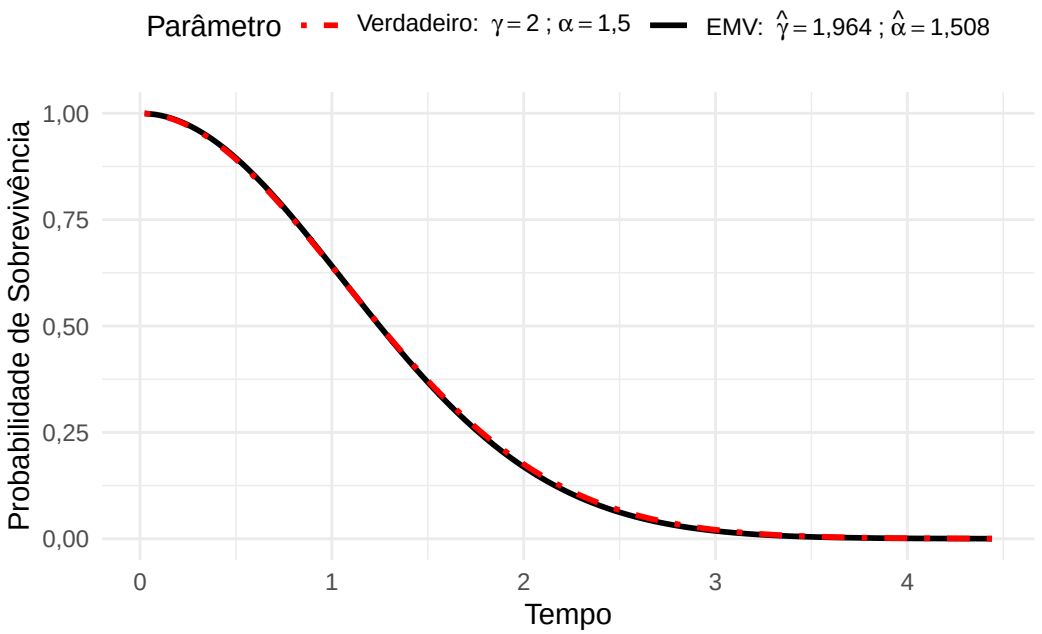
Assim como para distribuição exponencial, será feita uma comparação entre o real e estimado. Veja a Tabela 3.2 que mostra as dez observações, na ordem de simulação, e suas respectivas funções de sobrevivência, real e estimada.

Tabela 3.2: Comparação de Dez Observações entre o valor Real e o Estimado por Máxima Verossimilhança da Função de Sobrevivência segundo o Modelo Weibull.

Tempo	$S(t)$	$\widehat{S}_{EMV}(t)$
0,9081	0,6932	0,6912
0,9442	0,6729	0,6711
0,9776	0,6539	0,6524
0,8616	0,7190	0,7166
0,4288	0,9215	0,9189
0,4892	0,8991	0,8962
1,8050	0,2350	0,2408
0,4011	0,9310	0,9285
1,7208	0,2682	0,2735
1,3342	0,4533	0,4554

Também foi feita a comparação entre as duas curvas de sobrevivência, ilustradas na Figura 3.5.

Figura 3.5: Comparação das Curvas de Sobrevivência Real e Estimada por Máxima Verossimilhança segundo o Modelo Weibull.



3.3.3.3 Distribuição Log-normal

Para a estimação dos parâmetros do modelo log-normal deve-se, também, utilizar o método numérico de Newton-Raphson. Logo, utilizaremos o método para maximar a seguinte função:

$$L(\theta) = \prod_{i=1}^{n} \left[\frac{1}{t_i \sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{\ln(t_i) - \mu}{\sigma} \right)^2 \right\} \right]^{\delta_i} \left[\Phi \left(\frac{-\ln(t_i) + \mu}{\sigma} \right) \right]^{1-\delta_i}. \quad (3.15)$$

Foi simulada uma amostra oriunda de uma distribuição log-normal com parâmetro de locação $\mu = 0$ e parâmetro de escala $\sigma^2 = 1$ e, a partir dessa amostra, obteve-se a estimativa de máxima verossimilhança para o parâmetro $\theta = (\mu, \sigma^2)$. Veja a Tabela 3.3, que apresenta as dez observações com as funções de sobrevivência real e estimada.

```
$par
[1] 0,01217 1,01342

$value
[1] 1444

$counts
function gradient
      72       18

$convergence
[1] 0

$message
NULL

$hessian
          [,1]      [,2]
[1,] 9,737e+02 1,271e-04
[2,] 1,271e-04 1,947e+03
```

Tabela 3.3: Comparação de Dez Observações entre o valor Real e o Estimado por Máxima Verossimilhança da Função de Sobrevivência segundo o Modelo Log-normal.

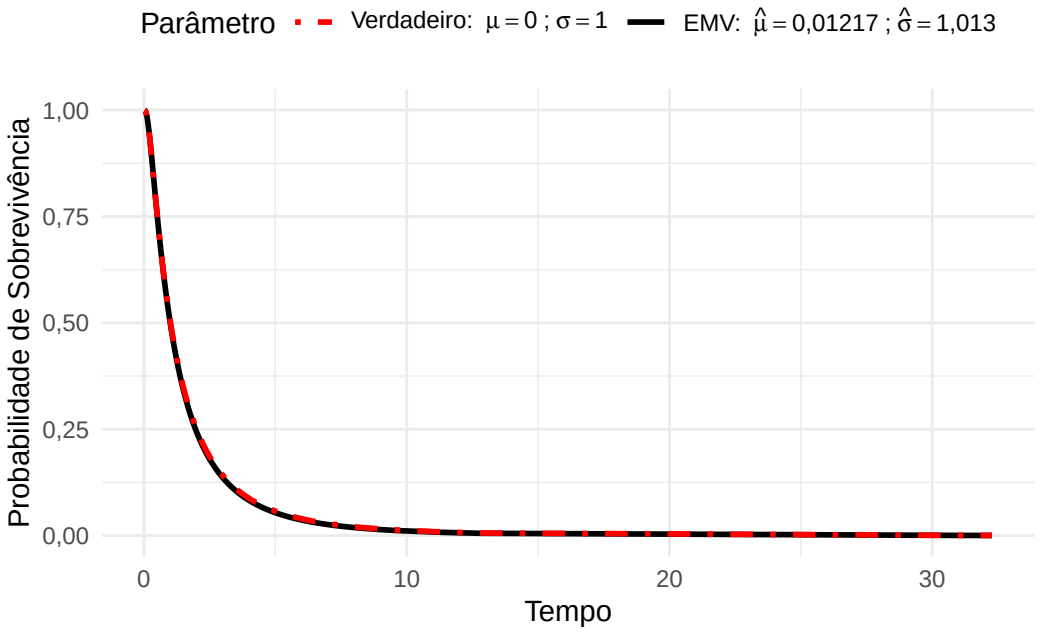
Tempo	$S(t)$	$\widehat{S}_{EMV}(t)$
1,6568	0,3068	0,3134
1,4857	0,3461	0,3525
4,1187	0,0785	0,0831
0,4856	0,7650	0,7657
0,5388	0,7318	0,7331
0,2096	0,9409	0,9399
1,1365	0,4491	0,4545
0,8547	0,5624	0,5663
0,2197	0,9352	0,9341

Tabela 3.3: Comparação de Dez Observações entre o valor Real e o Estimado por Máxima Verossimilhança da Função de Sobrevivência segundo o Modelo Log-normal.

Tempo	$S(t)$	$\widehat{S}_{EMV}(t)$
3,1950	0,1227	0,1284

Para uma visualização gráfica da otimização, foi criada a Figura 3.6.

Figura 3.6: Comparação das Curvas de Sobrevivência Real e Estimada por Máxima Verossimilhança segundo o Modelo Log-normal.



3.3.3.4 Distribuição Exponencial por Partes

Assim como para as distribuições Weibull e log-normal, será aplicado o método iterativo de Newton-Raphson. Porém, para a implementação do modelo exponencial por partes será usado o pacote `eha`. Esse pacote contém uma vasta quantidade de funções e implementações de modelos *Hazard Constant* (Risco Constante). Os parâmetros escolhidos para simulação dos dados foram: $\tau = (0,4; 1,2; 1,8)$ e $\lambda = (0,5; 1; 1,5; 2)$. Salientando que os parâmetros estimados serão apenas o vetor do parâmetro de taxas. Veja, abaixo, o ajuste do modelo através da maximização da seguinte função:

$$L(\theta) = \prod_{i=1}^{n} [...]^{\delta_i} [...]^{1-\delta_i} . \tag{3.16}$$

```
$par
[1] 0,5011 0,9831 1,4345 2,2094
```

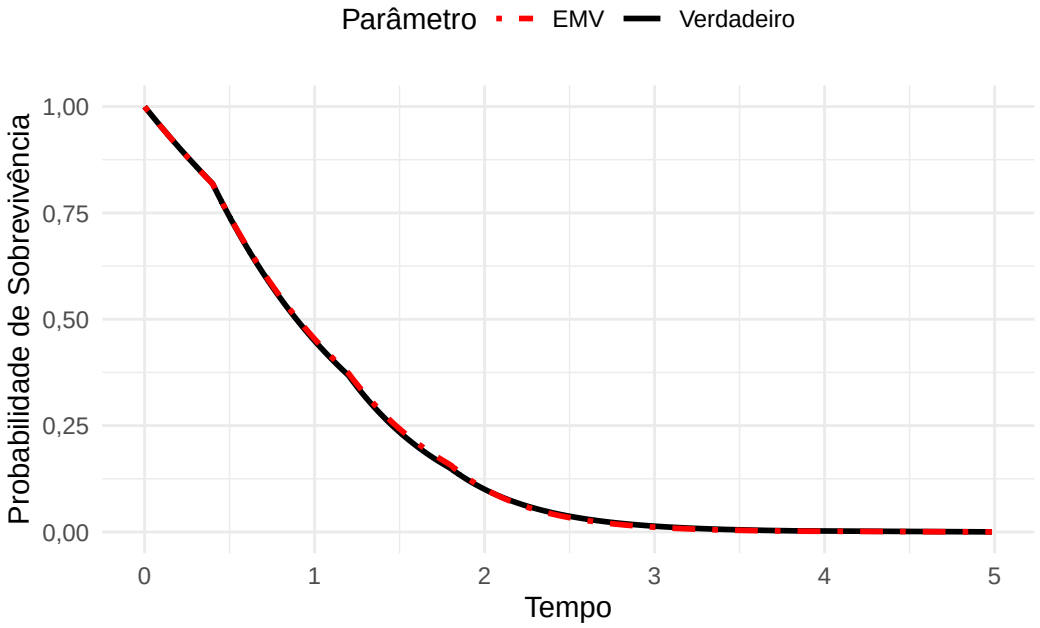
```
$value
[1] 926,2

$counts
function gradient
      52       13

$convergence
[1] 0

$message
NULL

$hessian
           [,1]        [,2]       [,3]       [,4]
[1,]  7,209e+02 -2,842e-08 0,000e+00 2,842e-08
[2,] -2,842e-08  4,574e+02 0,000e+00 2,842e-08
[3,]  0,000e+00  0,000e+00 1,040e+02 2,842e-08
[4,]  2,842e-08  2,842e-08 2,842e-08 3,339e+01
```

De forma semelhante aos demais modelos, será feita uma comparação da função de sobrevivência por meio da Tabela 3.4 e Figura 3.7. Mostradas a seguir.

Tabela 3.4: Comparação de Dez Observações entre o valor Real e o Estimado por Máxima Verossimilhança da Função de Sobrevivência segundo o Modelo Exponencial por Partes.

Tempo	$S(t)$	$\widehat{S}_{EMV}(t)$
1,3210	0,3068	0,3133
1,2783	0,3271	0,3331
1,2407	0,3461	0,3516
1,3796	0,2810	0,2881
2,1226	0,0785	0,0773
1,9968	0,1009	0,1020
0,4679	0,7650	0,7655
2,1868	0,0690	0,0671
0,5122	0,7318	0,7329
0,8039	0,5467	0,5502

Figura 3.7: Comparação das Curvas de Sobrevivência Real e Estimada por Máxima Verossimilhança segundo o Modelo Exponencial por Partes.



3.3.3.5 Distribuição Exponencial por Partes de Potência

Assim como para as distribuições anteriores, com excessão da distribuição exponencial clássica, o modelo exponencial por partes de potência requer a utilização do método numérico descrito na Seção 3.3.2. Para simulação dos dados foi usado como pontos de corte $\tau = (0,4; 1,2; 1,8)$, parâmetros de taxa $\lambda = (0,5; 1; 1,5; 2)$ e parâmetro de potência $\eta = 3/2$. As estimativas são obtidas maximizando a função:

$$L(\theta) = \prod_{i=1}^{n} [...]^{\delta_i} [...]^{1-\delta_i} . \tag{3.17}$$

```
$par
[1] 0,3907 0,9449 1,4424 1,9043 1,3467

$value
[1] 1011

$counts
function gradient
      70       20

$convergence
[1] 0
```

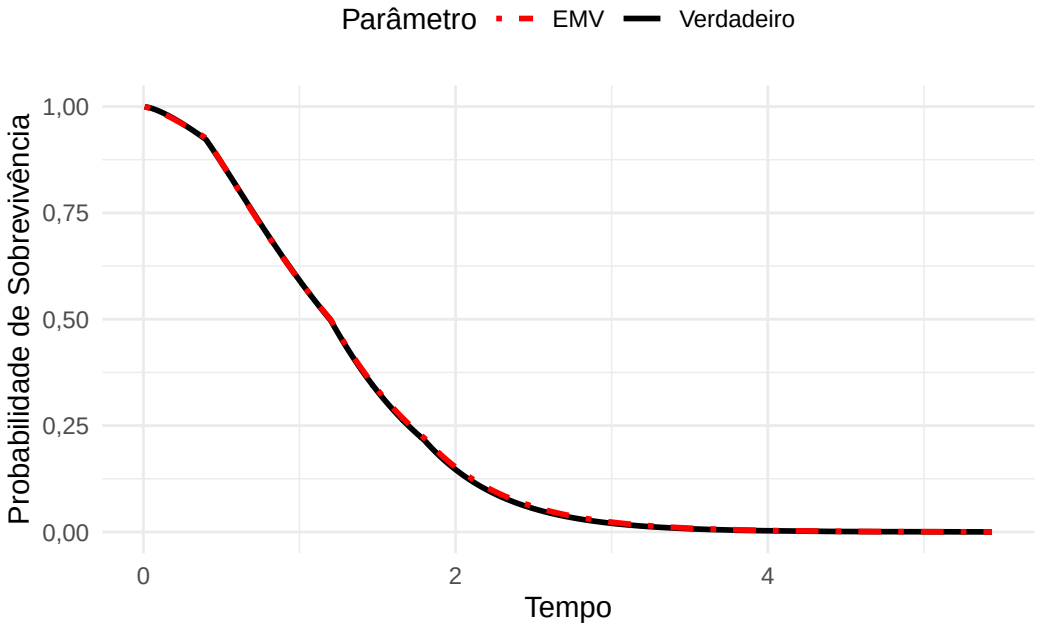
```
$message
NULL

$hessian
          [,1]     [,2]    [,3]    [,4]     [,5]
[1,]  838,0649   98,460   6,477  0,7767 -572,329
[2,]   98,4597  583,608  12,953  1,5534 -315,284
[3,]    6,4765   12,953 141,697  1,1651  -36,677
[4,]    0,7767    1,553   1,165 61,9192   -5,166
[5,] -572,3294 -315,284 -36,677 -5,1664  551,421
```

Veja a Tabela 3.5 e a Figura 3.8, que apresenta uma comparação da função de sobrevivência real e estimada.

Tabela 3.5: Comparação de Dez Observações entre o valor Real e o Estimado por Máxima Verossimilhança da Função de Sobrevivência segundo o Modelo Exponencial por Partes de Potência.

Tempo	$S(t)$	$\widehat{S}_{EMV}(t)$
0,8068	0,6932	0,6914
0,8439	0,6729	0,6713
0,8791	0,6539	0,6525
0,7604	0,7190	0,7171
0,4024	0,9215	0,9246
0,4443	0,8991	0,9009
1,7403	0,2350	0,2398
0,3684	0,9310	0,9332
1,6479	0,2682	0,2726
1,2696	0,4533	0,4555

Figura 3.8: Comparação das Curvas de Sobrevivência Real e Estimada por Máxima Verossimilhança segundo o Modelo Exponencial por Partes de Potência.



3.3.4 Aplicações Caso Haja Censura

Para exemplicação de ajustes dos modelos probabilísticos que foram propostos aqui à dados censurados, fixamos que os dados censurados como dados provenientes de uma distribuição exponencial com parâmetro de taxa $\alpha = 1$. Isto é, seja C uma variável aleatória, que representa os tempos de eventos censurados, tal quer $C \sim Exp(1)$.

Já a distribuição do tempo de evento, de fato, foi variada entre os modelos aqui propostos, logo, foram simulados tempos de evento segundo os modelos: $Weibull(2; 1, 5)$, $Lognormal(1, 2^2)$, $EP()$ e $EPP()$. Os tempos observados foram definidos de forma que, para cada unidade amostral, o tempo observado foi definido como a menor realização entre as duas distribuições em análise, ou seja:

$$t_i = min(T_i, C_i).$$

A censura ocorre quando o tempo de observação não corresponde ao tempo real de falha, ou seja, quando $C_i < T_i$. Nesse caso, o evento de interesse não foi completamente observado, sendo conhecido apenas que o verdadeiro tempo de falha excede o valor registrado. Essa característica, fundamental na análise de sobrevivência, requer métodos estatísticos específicos para garantir inferências adequadas a partir de dados censurados.

Sanando uma possível dúvida, que possa surgir da parte do leitor. Foi fixada para distribuição dos tempos de evento censurados a distribuição exponencial sem qualquer motivo em especial. Tendo em vista que a distribuição dos tempos de evento censurados não está sendo analisada. O obejto desta seção é apenas mostrar como os tempos censurados interferem na precisão das estimativas.

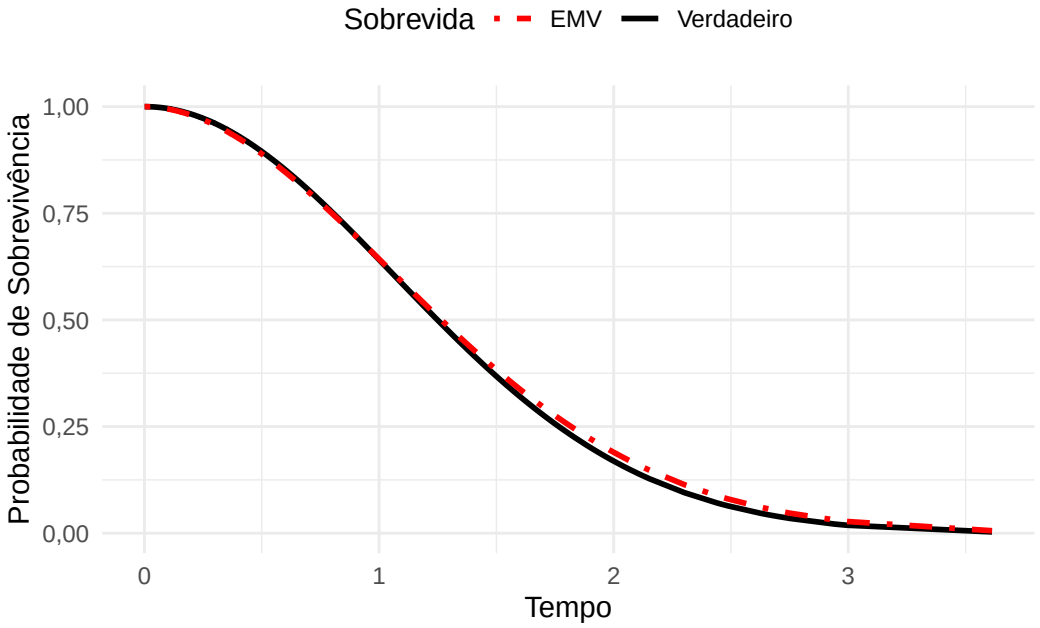
3.3.4.1 Modelo Weibull

Iniciando as exemplificações com a implementação do modelo Weibull. Veja a saída do ajuste para a distribuição $Weibull(2; 1, 5)$.

```
$par
[1] 1,911 1,533

$value
[1] 449,5

$counts
function gradient
      28        7

$convergence
[1] 0

$message
NULL

$hessian
       [,1]   [,2]
[1,] 187,59  78,67
[2,]  78,67 485,05
```

A seguir, uma visualização gráfica do ajuste ilustrada pela Figura 3.9.

Figura 3.9: Comparação das Curvas de Sobrevivência Real e Estimada por Máxima Verossimilhança segundo o Modelo Weibull para Dados Censurados.



3.3.4.2 Modelo Log-normal

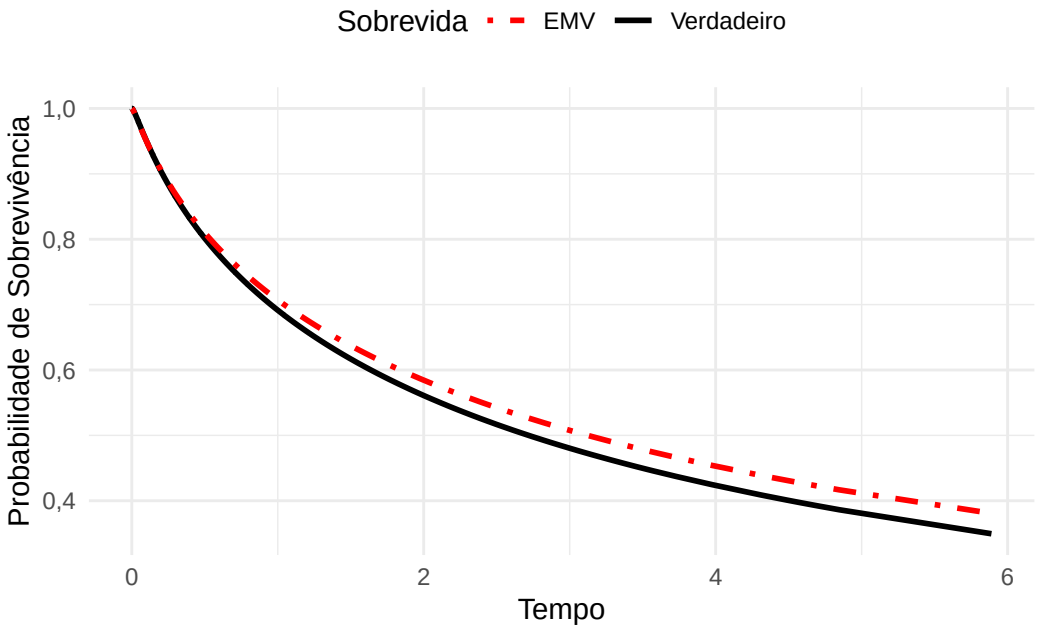
A segunda implementação feita, foi do modelo log-normal. Veja a saída do ajuste para a distribuição $Lognormal(1, 2^2)$.

```
$par
[1] 1,139 2,090

$value
[1] 480,5

$counts
function gradient
      78       35

$convergence
[1] 0

$message
NULL

$hessian
       [,1]   [,2]
[1,]  121,3 -109,5
```

```
[2,] -109,5  193,9
```

A seguir, uma visualização gráfica do ajuste ilustrada pela Figura 3.10.

Figura 3.10: Comparação de Dez Observações entre o valor Real e o Estimado por Máxima Verossimilhança da Função de Sobrevivência segundo o Modelo Log-normal para Dados Censurados.



3.3.4.3 Distribuição Exponencial por Partes

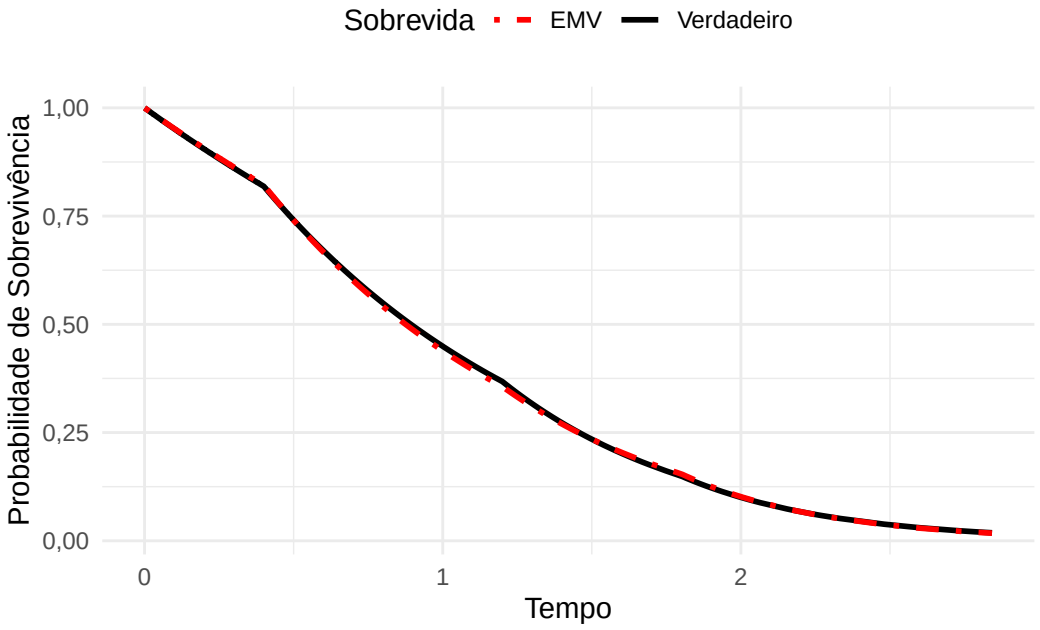
Fazendo uma implementação do modelo exponencial por partes para dados censurados. A saída do ajuste está logo abaixo para a distribuição $EP()$.

```
$par
[1] 0,4907 1,0450 1,3908 2,0843

$value
[1] 504,7

$counts
function gradient
      26       10

$convergence
[1] 0

$message
```

```
NULL

$hessian
           [,1]       [,2]       [,3]  [,4]
[1,]  6,063e+02 -1,421e-08  7,105e-09 0,000
[2,] -1,421e-08  2,042e+02 -1,421e-08 0,000
[3,]  7,105e-09 -1,421e-08  2,740e+01 0,000
[4,]  0,000e+00  0,000e+00  0,000e+00 5,294
```

A Figura 3.11 mostra o ajuste de forma gráfica fazendo uma comparação de curvas de sobrevivência.

Figura 3.11: Comparação das Curvas de Sobrevivência Real e Estimada por Máxima Verossimilhança segundo o Modelo Exponencial por Partes para Dados Censurados.



3.3.4.4 Distribuição Exponencial por Partes de Potência

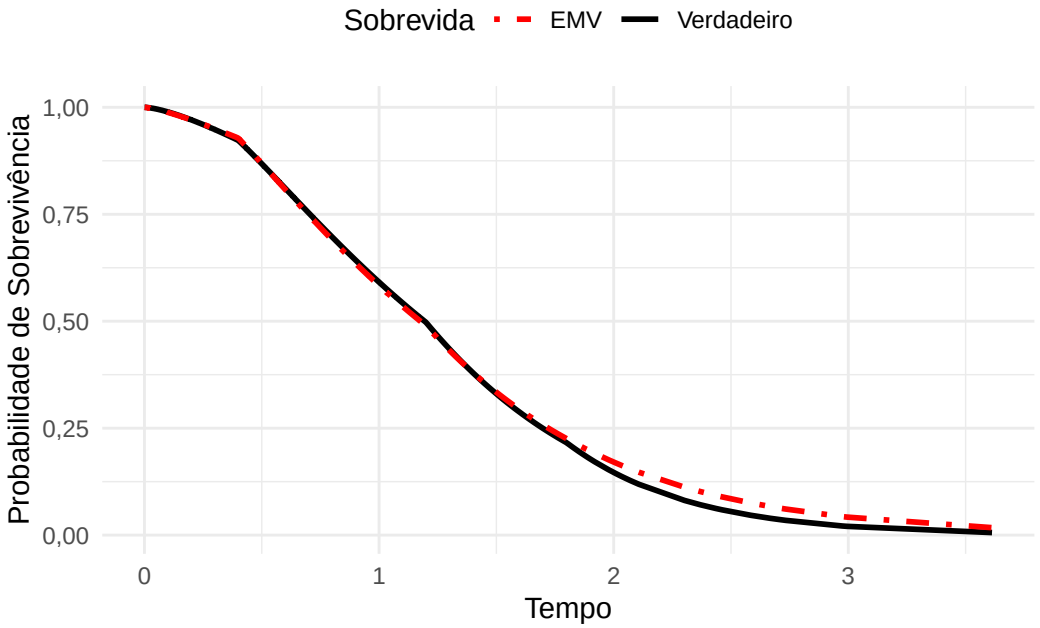
Por fim, é apresentado o ajuste para a distribuição $EPP()$.

```
$par
[1] 0,3853 0,9786 1,3719 1,4265 1,3525

$value
[1] 461

$counts
function gradient
```

```
      40       13

$convergence
[1] 0

$message
NULL

$hessian
          [,1]      [,2]    [,3]    [,4]      [,5]
[1,]  668,6728   51,1112  1,5572  0,1151 -450,8718
[2,]   51,1112  254,4160  3,1144  0,2303 -153,0120
[3,]    1,5572    3,1144 33,9231  0,1727   -8,6832
[4,]    0,1151    0,2303  0,1727  9,4803   -0,7445
[5,] -450,8718 -153,0120 -8,6832 -0,7445  397,0270
```

A seguir, tem-se a comparação das curvas de sobrevivência real e estimada, desenhada na Figura 3.12, para os dados simulados com censura.

Figura 3.12: Comparação das Curvas de Sobrevivência Real e Estimada por Máxima Verossimilhança segundo o Modelo Exponencial por Partes de Potência para Dados Censurados.



4 Modelos de Tempo de Vida Acelerado

4.1 Introdução

No capítulo anterior, foram apresentados modelos paramétricos para dados de sobrevivência. Entretanto, esses modelos não contemplam a inclusão de covariáveis na análise do tempo de sobrevivência. Neste capítulo, exploraremos esse método.

No modelo de regressão linear clássico, a relação entre a variável resposta Y e as covariáveis $\mathbf{x}'$ é aditiva, ou seja, mudanças nas covariáveis alteram Y de maneira linear. O modelo de regressão linear clássico é expresso como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon, \tag{4.1}$$

onde ε é a parte estocástica (erro) que segue uma distribuição $Normal(0, \sigma^2)$.

No entanto, em análise de sobrevivência, essa suposição não se sustenta, pois o efeito das covariáveis geralmente acelera ou retarda o tempo de falha, tornando necessária uma abordagem multiplicativa. Este modelo de regressão é chamado de Modelo de *Tempo de Vida Acelerado* (Accelerated Failure Time - AFT).

No modelo AFT, assume-se que o tempo de falha T é afetado por um fator de aceleração exponencial das covariáveis. Esse fator multiplicativo indica se o tempo até o evento será prolongado ou encurtado. Assim, o modelo é definido como:

$$T = \exp\{\mathbf{x}'\beta\}\varepsilon = \exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_p X_p\}\varepsilon, \tag{4.2}$$

onde ε é um termo de erro multiplicativo que captura a variabilidade não explicada pelas covariáveis. Aplicando a transformação logarítmica em T obtém-se a forma linearizável de Equação 4.2 que aproxima-se da Equação 4.1, de forma que

$$\ln[T] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_p X_p + v,$$

onde $v = \ln[\varepsilon]$ segue uma distribuição de valor extremo. Essa escolha para a distribuição dos erros decorre do fato de que os tempos de sobrevivência frequentemente apresentam forte assimetria à direita. Portanto, os erros não podem ser adequadamente representados por uma distribuição normal, sendo mais apropriado assumir distribuições como Log-normal, Weibull ou Exponencial.

Nos modelos AFT, a função de sobrevivência sofre um ajuste devido ao efeito das covariáveis, que podem acelerar ou retardar o tempo de falha. Assim, a função de sobrevivência condicional às covariáveis é expressa como:

$$S(t|x) = P(T > t/\exp\{\mathbf{x}'\beta\}). \tag{4.3}$$

Como o tempo de falha é ajustado pelo fator de aceleração, a função de risco também precisa ser reformulada para incorporar o efeito das covariáveis. A forma geral da função de risco em modelos AFT é dada por:

$$\lambda(t|\mathbf{x}) = \lambda_0(t)g(\mathbf{x}). \tag{4.4}$$

Nesta expressão, $\lambda_0(t)$, representa a função de risco basal, isto é, representa o risco no tempo t quando todas as covariáveis são iguais a zero, ou seja, na ausência de efeitos das covariáveis. Já o termo $g(\mathbf{x}) = \exp\{-\mathbf{x}'\beta\}$ age como um fator de ajuste, mensurando o impacto das covariáveis na taxa de falha.

4.2 Modelo Exponencial

Em modelos AFT, a função de sobrevivência à distribuição exponencial é expressa por:

$$S(t|x) = \exp\left\{-\alpha\left(\frac{t}{\exp\{\mathbf{x}'\beta\}}\right)\right\}. \tag{4.5}$$

Com função de risco dada por:

$$\lambda(t|\mathbf{x}) = \alpha\exp\{-\mathbf{x}'\beta\}. \tag{4.6}$$

4.3 Modelo Weibull

Para modelos AFT, baseados na distribuição Weibull, a função de sobrevivência é dada por:

$$S(t|x) = \exp\left\{-\left(\frac{t}{\alpha\exp\{-\mathbf{x}'\beta\}}\right)^{\gamma}\right\}. \tag{4.7}$$

Assim, pode-se escrever a função de risco da distribuição Weibull como:

$$\lambda(t|\mathbf{x}) = \frac{\gamma}{\alpha^{\gamma}}t^{\gamma-1}\exp\{-\mathbf{x}'\beta\} \tag{4.8}$$

4.4 Modelo Exponencial por Partes

[...]

4.5 Estimação de Parâmetros

Assim como no capítulo anterior, a estimação dos parâmetros será realizada pelo método de máxima verossimilhança. Recordando que a função de verossimilhança para dados censurados é expressa como:

$$L(\theta) = \prod_{i=1}^{n} \left[f(t_i|\mathbf{x})\right]^{\delta_i} \left[S(t_i|\mathbf{x})\right]^{1-\delta_i}$$
$$= \prod_{i=1}^{n} \left[\lambda(t_i|\mathbf{x})\right]^{\delta_i} S(t_i|\mathbf{x}),$$

onde:

- δ_i é a variável indicadora, assumindo 1 se t_i for um tempo de falha observado e 0 se for censurado;
- $f(t_i|\mathbf{x})$ representa a função densidade de probabilidade condicional;
- $S(t_i|\mathbf{x})$ é a função de sobrevivência condicional;
- $\lambda(t_i|\mathbf{x})$ corresponde à função de risco condicional.

O estimador de máxima verossimilhança (EMV) para θ é obtido maximizando a função de log-verossimilhança, dada por:

$$\ln L(\theta) = \sum_{i=1}^{n} \delta_i \ln \lambda(t_i|\mathbf{x}) + \ln S(t_i|\mathbf{x}).$$

Portanto, o parâmetro ou o conjunto de parâmetros θ que maximiza $\ln L(\theta)$ representa a melhor estimativa para a amostra observada, sendo obtido por métodos numéricos como o Newton-Raphson.

4.6 Implementação Computacional

4.6.1 Modelo Exponencial

4.6.1.1 Geração dos Dados

- **Funções de Geração:**

Pode-se simular dados de sobrevivência conforme um modelo AFT baseado na distribuição exponencial através da expressão:

$$T = -\frac{\exp\{\mathbf{x}'\beta\} \ln\left[1 - U\right]}{\alpha},$$

onde $U \sim Uniforme(0, 1)$. Sendo este o *Método da Tranformação da Inversa.*

- **Simulação dos dados:**

Veja a Tabela 4.1 que apresenta as dez primeiras observações simuladas.

Tabela 4.1: Dez primeiras observações Simuladas para Dados de Sobrevivência Censurados baseados no Modelo Exponencial.

Tempo	Delta	Constante	X1 ~ Bern(0,5)	X2 ~ Normal(0, 1)
0,5804	0	1	1	0,2627
0,8607	1	1	1	0,3934
0,1656	0	1	1	1,3079
0,6948	1	1	1	-0,2373
0,2539	1	1	1	-0,3461
0,0724	1	1	1	0,7248
0,0019	0	1	0	-0,4015
0,0147	1	1	1	-1,3736
1,6933	0	1	0	0,6389
0,9775	0	1	0	0,0040

A proporção de censura nos dados foi: 44,8%.

4.6.1.2 Ajuste do modelo ATF usando o pacote `survival`:

```
Call:
survreg(formula = Surv(times, delta) ~ x1 + x2, data = dados,
    dist = "exponential")
             Value Std. Error     z      p
(Intercept)  0,0688     0,0720  0,96   0,34
x1          -0,4648     0,0951 -4,89  1e-06
x2           0,4674     0,0476  9,82 <2e-16

Scale fixed at 1

Exponential distribution
Loglik(model)= -313,9   Loglik(intercept only)= -369,5
    Chisq= 111,2 on 2 degrees of freedom, p= 7e-25
Number of Newton-Raphson Iterations: 5
n= 1000
```

4.6.1.3 Usando a função `optim`

- **Implementando a função log-verossimilhança:**

A função log-verossimilhança, a ser maximizada, para um modelo AFT baseado na distribuição exponencial é dada por:

$$\ln L(\theta) = \sum_{i=1}^{n} \delta_i \ln \alpha \exp\{-\mathbf{x'}_i\beta\} + \ln \exp\left\{-\alpha\left(\frac{t_i}{\exp\{\mathbf{x'}_i\beta\}}\right)\right\}.$$

- **Maximizando:**

```
$par
[1]  1,5527  0,5088 -0,4647  0,4674

$value
[1] 313,9

$counts
function gradient
      39        9

$convergence
[1] 0

$message
NULL

$hessian
        [,1]   [,2]    [,3]    [,4]
[1,]  185,82 -288,5 -159,72   68,45
[2,] -288,53  448,0  248,00 -106,28
[3,] -159,72  248,0  248,00  -49,45
[4,]   68,45 -106,3  -49,45  467,08
```

4.6.2 Modelo Weibull

Pode-se simular dados de sobrevivência conforme um modelo AFT baseado na distribuição Weibull através da expressão:

$$T = -\alpha \exp\{\mathbf{x}'\beta\}(\ln[1-U])^{1/\gamma},$$

onde $U \sim Uniforme(0,1)$. Sendo este o *Método da Tranformação da Inversa*.

4.6.2.1 Geração dos Dados

- **Funções de Geração:**

- **Simulação dos dados:**

Veja a Tabela 4.2 que apresenta as dez primeiras observações simuladas.

Tabela 4.2: Dez primeiras observações Simuladas para Dados de Sobrevivência Censurados baseados no Modelo Weibull.

Tempo	Delta	Constante	X1 ~ Bern(0,5)	X2 ~ Normal(0, 1)
0,8706	0	1	1	0,2627
1,4803	0	1	1	0,3934
0,2484	0	1	1	1,3079
1,3282	0	1	1	-0,2373
0,8490	1	1	1	-0,3461
0,4201	0	1	1	0,7248
0,0028	0	1	0	-0,4015
0,1581	1	1	1	-1,3736
2,5399	0	1	0	0,6389
1,4662	0	1	0	0,0040

A proporção de censura nos dados foi: 27,5%.

4.6.2.2 Ajuste do modelo ATF usando o pacote `survival`:

```
Call:
survreg(formula = Surv(times, delta) ~ x1 + x2, data = dados,
    dist = "weibull")
              Value Std. Error      z       p
(Intercept)  0,8694     0,0524  16,58 < 2e-16
x1          -0,4534     0,0622  -7,29 3,2e-13
x2           0,4803     0,0316  15,21 < 2e-16
Log(scale)  -0,6866     0,0427 -16,07 < 2e-16

Scale= 0,503

Weibull distribution
Loglik(model)= -387,6   Loglik(intercept only)= -503,3
    Chisq= 231,3 on 2 degrees of freedom, p= 6e-51
Number of Newton-Raphson Iterations: 7
n= 1000
```

4.6.2.3 Usando a função `optim`

- **Implementando a função log-verossimilhança:**

A função log-verossimilhança, a ser maximizada, para um modelo AFT baseado na distribuição Weibull é dada por:

$$\ln L(\theta) = \sum_{i=1}^{n} \delta_i \ln \frac{\gamma}{\alpha^\gamma} t_i^{\gamma-1} \exp\{-\mathbf{x'}_i\beta\} + \ln \exp\left\{-\left(\frac{t_i}{\alpha \exp\{-\mathbf{x'}_i\beta\}}\right)^\gamma\right\}.$$

- **Maximizando:**

```
$par
[1]  1,9870  0,9900  0,8795 -0,4534  0,4803

$value
[1] 387,6

$counts
function gradient
      61       17

$convergence
[1] 0

$message
NULL

$hessian
       [,1]   [,2]   [,3]    [,4]    [,5]
[1,] 159,41  115,5  114,3   47,73   34,15
[2,] 115,48 1107,9 1096,7  646,08 -478,11
[3,] 114,32 1096,7 1085,7  639,59 -473,31
[4,]  47,73  646,1  639,6  639,59 -246,20
[5,]  34,15 -478,1 -473,3 -246,20 1265,04
```

4.6.3 Modelo Exponencial por Partes

Referêcias

Aalen, Odd O. 1978. «Nonparametric Inference for a Family of Counting Processes». *Annals of Statistics* 6 (4): 701–26. https://doi.org/10.1214/aos/1176344247.

Aalen, Odd O., e Søren Johansen. 1978. «An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations». *Scandinavian Journal of Statistics* 5 (3): 141–50.

Bohoris, G. A. 1994. «Comparison of the Cumulative-Hazard and Kaplan-Meier Estimators of the Survivor Function». *IEEE Transactions on Reliability* 43 (2): 230–32. https://doi.org/10.1109/24.293488.

Breslow, Norman, e John Crowley. 1974. «A Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship». *The Annals of Statistics* 2 (3): 437–53. https://doi.org/10.1214/aos/1176342705.

Bussab, Wilton de Oliveira, e Pedro Alberto Morettin. 2010. *Estatística Básica*. 6ª ed. São Paulo: Saraiva.

Colosimo, Enrico Antonio, e Suely Ruiz Giolo. 2006. *Análise de Sobrevivência Aplicada*. 1.ª ed. São Paulo, Brasil: Blucher.

Gehan, Edmund A. 1965. «A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples». *Biometrika* 52 (1-2): 203–24. https://doi.org/10.2307/2333825.

Kalbfleisch, John D., e Ross L. Prentice. 1980. *The Statistical Analysis of Failure Time Data*. Wiley Series em Probability e Mathematical Statistics. New York: Wiley.

Kaplan, Edward L., e Paul Meier. 1958. «Nonparametric Estimation from Incomplete Observations». *Journal of the American Statistical Association* 53 (282): 457–81. https://doi.org/10.1080/01621459.1958.10501452.

Klein, John P. 1991. «Small Sample Moments of Some Estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators». *Scandinavian Journal of Statistics* 18 (4): 333–40. https://doi.org/10.2307/4616203.

Latta, Robert B. 1981. «A Monte Carlo Study of Some Two-Sample Rank Tests with Censored Data». *Journal of the American Statistical Association* 76 (375): 713–19. https://doi.org/10.2307/2287572.

Lawless, J. F. 1982. *Statistical Models and Methods for Lifetime Data*. Wiley Series em Probability e Statistics. New York: John Wiley & Sons.

Lindsey, Jane C., e Louise M. Ryan. 1998. «Methods for Interval-Censored Data». *Statistics in Medicine* 17 (2): 219–38. https://doi.org/10.1002/(SICI)1097-0258(19980130)17:2%3C219::AID-SIM735%3E3.0.CO;2-D.

Mantel, Nathan. 1966. «Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration». *Cancer Chemotherapy Reports* 50 (3): 163–70.

Mantel, Nathan, e William Haenszel. 1959. «Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease». *Journal of the National Cancer Institute* 22 (4): 719–48.

Meier, Paul. 1975. «Estimation of a Survival Curve from Incomplete Data». *Journal of the American Statistical Association* 70 (351): 607–10. https://doi.org/10.1080/01621459.1975.10479872.

Nelson, Wayne. 1972. «Theory and Applications of Hazard Plotting for Censored Failure Data». *Technometrics* 14 (4): 945–66. https://doi.org/10.1080/00401706.1972.10488981.

Peto, Richard, e Julian Peto. 1972. «Asymptotically Efficient Rank Invariant Test Procedures». *Journal of the Royal Statistical Society: Series A (General)* 135 (2): 185–98. https://doi.org/10.2307/2344317.

Prentice, Ross L. 1978. «Linear Rank Tests with Right Censored Data». *Biometrika* 65 (1): 167–79. https://doi.org/10.2307/2335206.

Turnbull, Bruce W. 1974. «Nonparametric Estimation of a Survivorship Function with Doubly Censored Data». *Journal of the American Statistical Association* 69 (345): 169–73. https://doi.org/10.1080/01621459.1974.10480146.