

Universidade de Brasília
Departamento de Estatística

Modelos de Mistura AFT para Sobrevida em Insuficiência Cardíaca Baseados na Fração de Ejeção

Maxsuell Nina Da Silva

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Maxsuell Nina Da Silva

Modelos de Mistura AFT para Sobrevida em Insuficiência Cardíaca Baseados na Fração de Ejeção

Orientador(a): Peter Zörnig

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Dedico este trabalho em honra aos meus pais e tios, que sempre me incentivaram a estudar e a buscar os meus sonhos.

Com especial gratidão, dedico também ao professor Daniel de matemática, que acreditou em meu potencial durante o período escolar.

Agradecimentos

Primeiramente, agradeço a Deus pela concretização deste projeto e pelo cumprimento do seu propósito em minha vida. Estendo meu profundo reconhecimento e gratidão a todos os amigos, familiares e colegas de trabalho que ofereceram apoio e acompanhamento incondicional ao longo desta jornada.

Resumo

O presente estudo teve como objetivo analisar a influência da Fração de Ejeção do Ventrículo Esquerdo (FEVE) na sobrevida de pacientes com Insuficiência Cardíaca (IC). Foi conduzido um estudo experimental com 299 pacientes (105 mulheres e 194 homens) sintomáticos para IC, cujos dados foram coletados no Instituto de Cardiologia de Faisalabad e no Hospital Allied, localizados em Punjab, Paquistão, entre abril e dezembro de 2015. No inicio do estudo , foi aplicado métodos convêncionais como o estimador de Kaplan-Meier e o Modelo de Riscos Proposrcionais de Cox. Contudo, na análise convêncional de sobrevivência pressupõe que todos os individuos eventualmente experimentarão o evento de interesse, o que não condizia com o comportamento da amostra. A curva de Kaplan-Meier apresentou estabilização em uma proporção não zero, com censoramento itenso na cauda, sugerindo a existência de uma 'fração curada' ou de sobreviventes de longo prazo. Diante da violação dos pressupostos de riscos proporcionais e da natureza da grande quantidade de censuras, foi aplicado o modelo de mistura de tempo de vida acelerado (AFT) para identificar os subgrupos de risco. Foram testadas distribuições Weibull, Lognormal e Gumbel no Modelo AFT, e a seleção do melhor ajuste foi baseada nos critérios AIC, BIC e Index. Verificou-se que o Modelo AFT Gumbel apresentou o melhor ajuste aos dados. A conclusão foi que as classificações tradicionais de FEVE e NYHA não são suficientes para verificar de forma abrangente o risco de morte por insuficiência cardíaca ,como já era o esperado, e que a estratificação do risco por meio desses modelos pode auxiliar na identificação de subgrupos de pacientes com pior prognóstico, oferecendo subsídios para a tomada de decisão clínica.

Lista de Tabelas

1 Propriedades da Distribuição Exponencial 20

2 Propriedades da Distribuição de Weibull 21

3 Propriedades da Distribuição Log-Normal 21

4 Descrição das variáveis do conjunto de dados de pacientes com insuficiência cardíaca. 40

5 Primeiras seis observações do conjunto de dados analisado. 40

6 Estatísticas Descritivas das Variáveis Contínuas 41

7 Resumo das Variáveis Fatoriais (Categóricas) 41

8 Critério de Informação de Akaike (AIC) para diferentes distribuições ajustadas 57

9 Comparação dos modelos de mistura AFT ajustados 60

10 Resultado do Teste Log-Rank para os Clusters do Modelo gumbel K=2 . . 65

Lista de Figuras

1 Responsabilidade vs. FE (Gumbel AFT K=2) 61

2 Responsabilidade vs. FE (Weibull AFT K=2) 62

3 Responsabilidade vs. FE (Lognormal AFT K=2) 63

4 Curvas de Sobrevivência de Kaplan-Meier por Cluster (Modelo Gumbel AFT K=2) 64

Sumário

1 Introdução 8

2 Objetivos 9

3 Referencial Teórico 10

3.1 Insuficiência Cardíaca: Definição, Classificação e Prognóstico 10

3.2 Definição e Tipos de Insuficiência Cardíaca 10

3.3 Avaliação Funcional e Desafio Prognóstico 11

3.4 O Papel dos Registros Eletrônicos de Saúde (EHRs) no Prognóstico 11

4 Análise de Sobrevivência 12

4.1 Tempo de Falha 12

4.2 Censura 12

5 Representação e Funções Fundamentais na Análise de Sobrevivência . . 13

5.1 Representação dos Dados 13

5.2 Funções de Sobrevivência e Risco 14

5.2.1 Função de Sobrevivência 14

5.2.2 Função Taxa de Falha (Risco) 14

5.2.3 Função Taxa de Falha Acumulada 15

5.3 Tempo Médio e Vida Média Residual 15

5.4 Relações entre as Funções 16

5.5 Relações Adicionais 16

6 Técnicas Não-Paramétricas para Análise de Sobrevivência 17

6.1 O Estimador de Kaplan-Meier 17

6.2 Precisão e Intervalos de Confiança 18

6.3 O Modelo de Riscos Proporcionais de Cox 18

6.4 A Suposição de Proporcionalidade 19

6.5 Estimação dos Parâmetros e Verossimilhança Parcial 19

6.6 Diagnóstico do Modelo 20

6.7 Modelos Probabilísticos para o Tempo de Falha 20

6.8 Estimação dos Parâmetros dos Modelos 22

7 Método de Máxima Verossimilhança 22

7.1 Precisão das Estimativas e Intervalo de Confiança 23

7.2 Escolha do Modelo Probabilístico. 23

8 Estimação por Máxima Verossimilhança. 26

8.1 Matriz de Informação de Fisher. 26

8.2 Matriz de Covariância Assintótica 27

9 O Algoritmo Expectation-Maximization 28

9.1 Relação Fundamental de Probabilidade 28

9.2 Log-Verossimilhança em Termos dos Dados Completos 28

9.3 Aplicação da Expectativa 28

9.4 Definição da Função Q. 28

9.5 Definição da Função H. 29

9.6 Estimativa do Intercepto e Robustez 31

9.7 Inferência por Máxima Verossimilhança 31

10 Modelo de Mistura AFT 32

10.1 Incidência via Regressão Logística. 33

10.2 Função de Log-Verossimilhança 33

10.3 Verossimilhança Completa 34

11 Estimação via Algoritmo EM 34

11.1 Verossimilhança Completa 34

11.2 Passo E (Expectation) 35

11.3 Passo M (Maximization) 35

12 Metodologia 37

12.1 Fluxo de Modelagem Estatística. 37

12.2 Avaliação e Vantagem Metodológica 38

12.3 Conjunto de dados 39

12.4 Descrição das variáveis do conjunto de dados 39

13 Aplicação do estudo 41

13.1 Análise Descritiva das Variáveis Contínuas 42

13.2 Análise da variável principal do estudo - FRAÇÃO DE EJEÇÃO 43

13.3 Justificativa da Bimodalidade 44

13.4 Análise do Boxplot da Fração de Ejeção 44

13.5 Análise Não Paramétrica de Sobrevivência 45

13.6 Descrição do Boxplot da Fração de Ejeção (FE) 46

13.7 Criação do modelo de Cox 54

13.8 Análise Paramétrica de Sobrevivência 56

13.9 Modelos de Mistura AFT com Diferentes Distribuições 58

13.9.1 Modelo de Mistura Weibull AFT 59

13.9.2 Modelo de Mistura Gumbel AFT 59

13.9.3 Modelo de Mistura Lognormal AFT 59

13.9.4 Síntese Comparativa 60

14 Análise dos gráficos 61

14.1 Análise Visual da Separação dos Clusters (Responsabilidade) 61

14.2 Modelo de Mistura Gumbel AFT (K=2) 61

14.3 Modelo de Mistura Weibull AFT (K=2) 62

14.4 Modelo de Mistura Lognormal AFT (K=2) 63

14.5 Conclusão Comparativa Visual 63

15 Conclusão 63

15.1 Validação Visual por Kaplan-Meier 64

15.2 Validação Estatística por Teste Log-Rank 64

15.3 Conclusão da Aplicação 65

16 Apêndice 65

1 Introdução

A classificação tradicional da Insuficiência Cardíaca (IC), baseada na Fração de Ejeção do Ventrículo Esquerdo (FEVE), é amplamente utilizada na prática clínica como um indicador da gravidade da doença. Entretanto, por ser uma classificação, ela pode conduzir a uma estratificação errada, pacientes com FEVE preservada podem apresentar maior mortalidade do que aqueles com FEVE reduzida, dependendo de seus perfis e históricos clínicos. O problema reside na heterogeneidade dos pacientes, cujos perfis distintos podem ser mascarados quando agrupados apenas pelo tipo de FEVE.

Modelos de sobrevivência clássicos, como o estimador de Kaplan-Meier ou o modelo de riscos proporcionais de Cox, assumem que os indivíduos dentro de um grupo são estatisticamente homogêneos. Essa premissa pode ocultar riscos extremos, dificultando a identificação de pacientes que demandam maior atenção e prejudicando a interpretação clínica.

A ideia central deste trabalho é investigar, por meio da variável FEVE, a existência de subgrupos de alto e baixo risco de morte por IC, utilizando os valores de cada observação em vez de categorias predefinidas. Dado que os dados clínicos frequentemente apresentam censura aleatória e um volume expressivo de dados faltantes, optou-se pela utilização do modelo de Tempo de Falha Acelerado (AFT) com o incremento de misturas pela provável existência de subgrupos na amostra. Este modelo é uma alternativa paramétrica robusta, cujos parâmetros foram estimados por Máxima Verossimilhança através do algoritmo Expectation-Maximization (EM).

Nessa análise comparou-se o comportamento da variável FEVE em modelos não-paramétricos e paramétricos, testando-se as distribuições Weibull, Log-normal, Gumbel e Exponencial. A seleção do melhor ajuste foi baseada nos critérios de informação AIC e BIC. No modelo de mistura AFT, a seleção de covariáveis partiu do modelo de Cox, utilizando o método de exclusão sucessiva até a obtenção do modelo mais parcimonioso (apenas com a FEVE), que apresentou os menores valores de AIC e BIC. Por fim, foram geradas curvas de Kaplan-Meier para os k grupos identificados, permitindo a distinção entre perfis de alto e baixo risco. Esta abordagem visa refinar a análise clínica e oferecer uma visão mais precisa da evolução da doença.

2 Objetivos

Esta pesquisa tem como propósito central investigar sistematicamente a heterogeneidade latente não observada nos padrões de sobrevida de pacientes diagnosticados com insuficiência cardíaca, mediante a aplicação avançada de Modelo de Mistura AFT especificamente adaptados para lidar com dados de sobrevivência censurados. O estudo busca superar as limitações inerentes às classificações tradicionais baseadas exclusivamente na Fração de Ejeção do Ventrículo Esquerdo (FEVE), desenvolvendo uma abordagem estratificadora mais refinada que capture subgrupos clinicamente relevantes com trajetórias de risco distintas que permanecem ocultas sob os paradigmas convencionais de avaliação.A fundamentação metodológica deste trabalho apoia-se na aplicação de modelos de regressão do tipo Tempo de Falha Acelerado (AFT) dentro de uma estrutura de mistura, utilizando o algoritmo Expectation-Maximization para estimação robusta dos parâmetros latentes, com o fim último de proporcionar uma ferramenta de estratificação de risco mais precisa e clinicamente útil para orientar condutas terapêuticas personalizadas nesta amostra de pacientes.

3 Referencial Teórico

3.1 Insuficiência Cardíaca: Definição, Classificação e Prognóstico

As Doenças Cardiovasculares (DCVs) representam um grupo heterogêneo de patologias que afetam o coração e os vasos sanguíneos. Dentre as DCVs mais prevalentes, destacam-se a Doença Arterial Coronariana (como o Infarto Agudo do Miocárdio), as Doenças Cerebrovasculares (como os Acidentes Vasculares Cerebrais) e a Insuficiência Cardíaca (IC) (DiretrizIC2018).

3.2 Definição e Tipos de Insuficiência Cardíaca

A Insuficiência Cardíaca é uma síndrome clínica complexa resultante da incapacidade do coração de bombear sangue em quantidade suficiente para atender às demandas metabólicas do corpo, ou de fazê-lo apenas com elevadas pressões de enchimento.

A classificação da IC baseia-se primariamente na Fração de Ejeção do Ventrículo Esquerdo (FEVE), que é a proporção de sangue ejetada pelo coração a cada batimento, expressa em porcentagem. Valores normais de FEVE situam-se tipicamente entre 50% e 75% (DiretrizIC2018). Os dois principais tipos são:

- **Insuficiência Cardíaca com Fração de Ejeção Reduzida (ICFEr ou HFrEF - *Heart Failure with Reduced Ejection Fraction*)**: Caracterizada por uma FEVE inferior a 40%. É classicamente conhecida como Insuficiência Cardíaca Sistólica ou disfunção ventricular esquerda.
- **Insuficiência Cardíaca com Fração de Ejeção Preservada (ICFEp ou HFpEF - *Heart Failure with Preserved Ejection Fraction*)**: Também denominada Insuficiência Cardíaca Diastólica. Neste tipo, a FEVE é $\geq 50\%$. A disfunção primária não é na contração, mas sim no relaxamento adequado do ventrículo esquerdo, prejudicando o seu enchimento e, consequentemente, o débito cardíaco.

Existe também uma categoria intermediária, a Insuficiência Cardíaca com Fração de Ejeção na Faixa Intermediária (ICFEnI ou HFmrEF - *Heart Failure with Mid-range Ejection Fraction*), com FEVE entre 40% e 49% (DiretrizIC2018).

3.3 Avaliação Funcional e Desafio Prognóstico

Para a avaliação clínica da gravidade e da progressão da doença, os médicos frequentemente utilizam a Classificação Funcional da *New York Heart Association* (NYHA), que categoriza a limitação dos pacientes em atividades diárias, conforme detalhado a seguir (DiretrizIC2018)

- **Classe I**: O paciente não tem nenhuma limitação para atividades físicas cotidianas.
- **Classe II**: O paciente apresenta sintomas leves ao realizar uma atividade física; sintomas (fadiga, dispneia) manifestam-se apenas com esforço intenso.
- **Classe III**: O paciente possui sintomas realizando atividades que exigem pouco esforço.
- **Classe IV**: O paciente possui sintomas realizando atividades físicas e até em repouso.

Apesar de sua ampla utilização, a classificação da NYHA apresenta limitações, especialmente no que tange à predição precisa de características clínicas fundamentais, como a tolerância ao exercício.

Dada a importância vital da função cardíaca, a predição do desfecho da Insuficiência Cardíaca é uma prioridade médica. Contudo, mesmo com os avanços significativos na pesquisa clínica, a previsão do prognóstico e da sobrevida dos pacientes com IC permanece um desafio com precisão frequentemente limitada (Chicco e Jurman, 2020).

3.4 O Papel dos Registros Eletrônicos de Saúde (EHRs) no Prognóstico

Neste contexto, os Registros Eletrônicos de Saúde (EHRs) emergem como uma fonte de dados de extrema relevância, capazes de fornecer uma visão longitudinal e detalhada do histórico clínico do paciente. Os EHRs podem potencialmente revelar correlações não óbvias entre diversos dados clínicos, laboratoriais e demográficos.

A aplicação de modelos analíticos avançados a esses registros médicos representa uma proposta poderosa. Tal abordagem não apenas visa aprimorar a capacidade de prever a sobrevida de pacientes com Insuficiência Cardíaca, mas também a identificar quais fatores clínicos são os mais relevantes e influentes no prognóstico.

Portanto, modelar a sobrevida na IC permanece um desafio complexo, tanto em termos de obtenção de alta precisão preditiva quanto na interpretação inequívoca dos múltiplos fatores que determinam o prognóstico do paciente (Chicco e Jurman, 2020).

4 Análise de Sobrevivência

Na análise de sobrevivência, a variável resposta é o tempo até a ocorrência de um evento de interesse. Um conjunto de dados de sobrevivência é caracterizado por dois componentes principais: os tempos de falha e as censuras.(PESCIM, R. R. Análise de Sobrevivência. [Introdução a análise de Sobrevivência no R]. YouTube. Londrina, 2024).

4.1 Tempo de Falha

A definição do tempo de falha é um elemento crucial. A análise requer um ponto de partida bem definido (origem), a partir do qual os indivíduos são acompanhados, e uma escala de medida. Embora a escala mais comum seja o tempo cronológico (dias, meses, anos), outras métricas podem ser utilizadas, como o número de ciclos ou a quilometragem em estudos de engenharia.

O evento de interesse, frequentemente denominado "falha", deve ser definido de maneira clara e inequívoca. Em muitos cenários, múltiplas causas podem levar à ocorrência da falha. Situações em que diferentes causas de falha competem entre si são conhecidas na literatura como riscos competitivos(Colosssimo, 2024).

4.2 Censura

Devido à longa duração de muitos estudos de acompanhamento, é comum que nem todos os indivíduos sejam observados até a ocorrência do evento de interesse. Pacientes podem abandonar o estudo, falecer por outras causas não relacionadas ao evento, ou o estudo pode ser encerrado antes que todos tenham apresentado o desfecho. Isso resulta em observações incompletas, conhecidas como dados censurados. A exclusão desses dados pode introduzir vieses significativos na análise.

Existem diferentes tipos de censura:

- **Censura do Tipo I:** Ocorre quando o estudo é encerrado após um período de tempo pré-estabelecido. Todos os indivíduos que não apresentaram o evento até essa data são considerados censurados.

- **Censura do Tipo II:** O estudo termina quando um número pré-definido de indivíduos apresenta o evento de interesse.
- **Censura Aleatória:** É o tipo mais frequente na prática clínica. Ocorre quando um indivíduo é retirado do estudo por motivos diversos (perda de seguimento, decisão pessoal, etc.) antes da ocorrência da falha.

Formalmente, a censura aleatória pode ser representada da seguinte forma: Seja T uma variável aleatória não negativa representando o tempo de falha de um indivíduo, e C uma variável aleatória não negativa representando o tempo de censura. Assume-se que T e C são independentes. O tempo observado para o indivíduo é t, definido como:

$$t = \min(T, C).$$

Adicionalmente, os mecanismos de censura podem ser classificados como:

- **Censura à Direita:** Ocorre quando o tempo de falha real é maior que o tempo de observação registrado. Sabemos que o indivíduo sobreviveu até um certo ponto, mas não sabemos o momento exato do evento. Este é o tipo mais comum de censura.
- **Censura à Esquerda:** Acontece quando o tempo de falha é menor que o tempo de observação. Sabemos que o evento ocorreu antes de um determinado momento, mas não sabemos exatamente quando.
- **Censura Intervalar:** Ocorre quando o evento de interesse é conhecido por ter acontecido dentro de um intervalo de tempo específico. Se T_i é o tempo de falha para o indivíduo i, então sabemos que $T_i \in (L_i, U_i]$, onde L_i e U_i são os limites inferior e superior do intervalo, respectivamente(Colosimo, 2024).

5 Representação e Funções Fundamentais na Análise de Sobrevivência

5.1 Representação dos Dados

Na Análise de Sobrevivência, o i-ésimo indivíduo ($i = 1, \ldots, n$) é usualmente representado pelo par (t_i, d_i), em que t_i é o tempo de observação (podendo ser tempo de

falha ou de censura) e d_i é a variável indicadora de evento (falha ou censura):

$$d_i = \begin{cases} 1 & \text{se } t_i \text{ é um tempo de falha (evento de interesse)} \\ 0 & \text{se } t_i \text{ é um tempo de censura (observação incompleta)} \end{cases}$$

No contexto de um estudo que inclui covariáveis, os dados ficam representados pelo trio $(t_i, d_i, \mathbf{x}_i)$, em que $\mathbf{x}_i$ é o vetor de covariáveis medidas no indivíduo i (por exemplo, $\mathbf{x}_i = (\text{sexo}, \text{idade}, \text{tratamento recebido})$)(Colosimo, 2024).

5.2 Funções de Sobrevivência e Risco

A Análise de Sobrevivência se concentra no estudo do tempo até a ocorrência de um evento, o qual é modelado por uma variável aleatória não-negativa T. As duas funções principais que especificam a distribuição de T são a função de sobrevivência e a função de taxa de falha (ou risco).

5.2.1 Função de Sobrevivência

A função de sobrevivência, denotada por $S(t)$, é a probabilidade de um indivíduo sobreviver além de um certo tempo t. Em termos probabilísticos, é definida como(Colosimo, 2024):

$$S(t) = P(T \geq t).$$

É uma função decrescente de t, com $S(0) = 1$ e $\lim_{t\to\infty} S(t) = 0$.

Por consequência, a função de distribuição acumulada $F(t)$ (a probabilidade do evento ocorrer antes ou em t) é dada por

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

5.2.2 Função Taxa de Falha (Risco)

A função taxa de falha (ou função de risco), denotada por $\lambda(t)$, representa a taxa instantânea de falha no tempo t, condicional ao indivíduo ter sobrevivido até aquele tempo. Matematicamente, ela é definida como

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

Esta função é essencial, pois descreve a "propensão"instantânea de falha ao longo do tempo. No contexto de densidade de probabilidade $f(t)$ e função de sobrevivência $S(t)$, temos a relação

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

5.2.3 Função Taxa de Falha Acumulada

A função taxa de falha acumulada, denotada por $\Lambda(t)$, é a integral da função de risco $\lambda(u)$ de 0 até t:

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

Embora não tenha uma interpretação probabilística direta como $S(t)$, a função $\Lambda(t)$ é fundamental na teoria de sobrevivência, estando intimamente ligada à função de sobrevivência pela relação

$$S(t) = e^{-\Lambda(t)}.$$

Essa relação é crucial para a estimação não paramétrica e para a construção de modelos de risco proporcional(Colosimo, 2024).

5.3 Tempo Médio e Vida Média Residual

Em análise de sobrevivência, duas quantidades de interesse bastante utilizadas são o tempo médio de vida e a vida média residual.

O tempo médio de vida corresponde à área sob a curva da função de sobrevivência $S(t)$. Assim, ele é dado por:

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

Já a vida média residual é definida de forma condicional a um tempo de vida t. Ou seja, para um indivíduo que já sobreviveu até o tempo t, essa quantidade representa o tempo médio de vida restante. Matematicamente, ela é calculada como a área sob a curva de sobrevivência à direita do tempo t, dividida pela probabilidade de sobreviver até t, isto é:

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

em que $f(\cdot)$ é a função densidade de probabilidade associada à variável aleatória T(Colosimo, 2024).

5.4 Relações entre as Funções

As funções $f(t)$, $S(t)$, $F(t)$ e $\lambda(t)$ (função risco) estão matematicamente relacionadas entre si. Essas relações são fundamentais na teoria da sobrevivência e podem ser expressas como:

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

A função *acumulada de risco*, denotada por $\Lambda(t)$, é definida como

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

e, consequentemente, a função de sobrevivência pode ser reescrita como

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

Essas expressões mostram que o conhecimento de uma dessas funções (por exemplo, $S(t)$) é suficiente para determinar todas as demais — $F(t)$, $f(t)$, $\lambda(t)$ e $\Lambda(t)$(Colosimo, 2024).

5.5 Relações Adicionais

Outras relações importantes entre a função de sobrevivência e a vida média residual são dadas por(Colosimo, 2024)

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

e

$$\lambda(t) = \left(\frac{d\,\mathrm{vmr}(t)}{dt} + 1\right) \Big/ \mathrm{vmr}(t).$$

6 Técnicas Não-Paramétricas para Análise de Sobrevivência

Na análise de dados de tempo até a falha, é comum a ocorrência de **censuras**. Uma observação censurada indica que o tempo exato até o evento (falha) não é conhecido, sabendo-se apenas que ele é maior do que o tempo registrado no estudo.

Para estimar a função de sobrevivência $S(t)$ a partir desses dados, utilizam-se técnicas estatísticas não paramétricas. As mais comuns são o estimador de Kaplan-Meier e o estimador de Nelson-Aalen(Colosimo, 2024).

6.1 O Estimador de Kaplan-Meier

O estimador de Kaplan-Meier, também conhecido como estimador limite-produto, é formalmente definido considerando as seguintes preliminares:

- $t_1 < t_2 < \cdots < t_k$: os k tempos distintos e ordenados em que ocorrem falhas.
- d_j: o número de falhas (eventos) no tempo t_j.
- n_j: o número de indivíduos sob risco no tempo t_j, ou seja, aqueles que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

A expressão geral do estimador de Kaplan-Meier é, então, definida como:

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

Este estimador possui propriedades estatísticas importantes:

i) É um estimador não viciado para amostras grandes.

ii) É fracamente consistente.

iii) Converge assintoticamente para um processo Gaussiano (distribuição Normal),Rasmussen e Williams (2006).

iv) É o estimador de máxima verossimilhança (MLE) de $S(t)$.

6.2 Precisão e Intervalos de Confiança

Para construir intervalos de confiança e realizar testes de hipóteses para $S(t)$, é fundamental avaliar a precisão do estimador $\hat{S}(t)$. Como todo estimador, ele está sujeito a variações amostrais, que são descritas por sua variância.

A expressão para a variância assintótica estimada de $\hat{S}(t)$ é conhecida como **fórmula de Greenwood**:

$$\widehat{Var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}.$$

Graças à propriedade de $\hat{S}(t)$ ter distribuição assintótica Normal (propriedade iii), podemos construir um intervalo de $(1-\alpha)\%$ de confiança para $S(t)$ da seguinte forma:

$$\hat{S}(t) \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{S}(t))}.$$

Em que $z_{\alpha/2}$ é o quantil superior $\alpha/2$ da distribuição Normal padrão.

É importante notar que este método pode, por vezes, gerar limites de intervalo inferiores a 0 ou superiores a 1. Para evitar esse problema, utilizam-se transformações (como a transformação log-log) que garantem que os limites do intervalo de confiança permaneçam dentro do intervalo $[0, 1]$(Colosimo, 2024).

6.3 O Modelo de Riscos Proporcionais de Cox

O Modelo de Riscos Proporcionais, também chamado modelo de taxas de falhas proporcionais, uma abordagem predominante na modelagem de dados clínicos e em análises de sobrevivência. A ampla preferência por este método deve-se, à sua flexibilidade e à capacidade de incorporar covariáveis dependentes do tempo de sobrevida.

A formulação do modelo pressupõe que os tempos de falha t_i $(i = 1, ..., n)$ são independentes. Matematicamente, a função de taxa de falha (ou risco), denotada por $\lambda(t)$, é expressa pela Equação.

$$\lambda(t) = \lambda_0(t) \exp\{\beta_1 x_1 + ... + \beta_p x_p\}.$$

Nesta expressão, $\lambda_0(t)$ representa uma função não-negativa do tempo, deno-

minada função de risco base (ou componente não paramétrico).O termo exponencial $\exp\{x'\beta\}$ é o componente de interesse principal, uma vez que os coeficientes β mensuram os efeitos das covariáveis explicativas sobre a função de taxa de falha(COLOSIMO, E. A. Métodos Estatísticos Avançados em Epidemiologia).

6.4 A Suposição de Proporcionalidade

A característica central deste modelo, que lhe confere o nome e facilita sua interpretação, é a propriedade das taxas de falhas proporcionais. O conceito estabelece que a razão entre as taxas de falha de dois indivíduos distintos, i e j, independe do tempo.

Essa relação é demonstrada matematicamente pela Equação

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\exp\{x_i'\beta\}}{\lambda_0(t)\exp\{x_j'\beta\}} = \exp\{x_i'\beta - x_j'\beta\}.$$

Nesse contexto, o efeito das covariáveis é interpretado como um fator de aceleração ou desaceleração da função de risco. Ao comparar dois indivíduos que possuem os mesmos valores para todas as covariáveis, exceto para a l-ésima variável, a razão das taxas de falha é dada por $\exp\{\beta_l(x_{il} - x_{jl})\}$.Este valor é interpretado clinicamente como a Razão de Riscos(COLOSIMO, E. A. Métodos Estatísticos Avançados em Epidemiologia).

6.5 Estimação dos Parâmetros e Verossimilhança Parcial

Para a realização de inferências estatísticas e estimação dos parâmetros, utiliza-se o estimador de máxima verossimilhança parcial, proposto por Cox (1975).Este método permite estimar os coeficientes β sem a necessidade de especificar o componente não paramétrico $\lambda_0(t)$.

A função de verossimilhança parcial $L(\beta)$ é constituída pelo produto das probabilidades condicionais de falha para todos os indivíduos da amostra, conforme a Equação

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\{x_i'\beta\}}{\sum_{j\in R(t)}\exp\{x_j'\beta\}} = \prod_{i=1}^{n}\left(\frac{\exp\{x_i'\beta\}}{\sum_{j\in R(t)}\exp\{x_j'\beta\}}\right)^{\delta_i}.$$

Na equação acima, δ_i representa o indicador de falha (evento).Os valores de β que maximizam a função $L(\beta)$ constituem os estimadores de máxima verossimilhança parcial.

6.6 Diagnóstico do Modelo

A adequação do modelo de Cox não é universal; a violação de sua suposição básica — a proporcionalidade dos riscos — pode resultar em vieses sérios na estimação dos coeficientes.Portanto, a avaliação da proporcionalidade das taxas de falha é uma etapa crítica.

Para diagnosticar a qualidade do ajuste, recomenda-se o uso de métodos gráficos, como o gráfico dos resíduos padronizados (Schoenfeld), que servem para observar a proporcionalidade das covariáveis, bem como a aplicação de testes de hipótese globais para a proporcionalidade de riscos(COLOSIMO, E. A. Métodos Estatísticos Avançados em Epidemiologia).

6.7 Modelos Probabilísticos para o Tempo de Falha

A modelagem paramétrica do tempo de falha é uma etapa essencial na análise de sobrevivência. Embora existam diversas distribuições de probabilidade aplicáveis, as distribuições Exponencial, Weibull e Log-Normal se destacam pela sua comprovada adequação em várias situações práticas. Cada distribuição possui propriedades e parâmetros distintos que influenciam na caracterização do tempo de vida (Colosimo, 2024).

1. Distribuição Exponencial

É um dos modelos mais simples, apresentando apenas um parâmetro ($\lambda > 0$, taxa de falha). Sua característica principal é a função de taxa de falha constante ($h(t) = \lambda$), refletindo a propriedade de falta de memória, em que a probabilidade de um evento ocorrer em um intervalo de tempo futuro é independente do tempo já vivido.

Tabela 1: Propriedades da Distribuição Exponencial

Propriedade	**Fórmula**
Função Densidade de Probabilidade ($f(t)$)	$f(t) = \lambda e^{-\lambda t}, \quad t \geq 0, \lambda > 0$
Função de Sobrevivência ($S(t)$)	$S(t) = e^{-\lambda t}$
Função Taxa de Falha ($h(t)$)	$h(t) = \lambda$
Média ($E[T]$)	$E[T] = 1/\lambda$
Variância (Var[T])	Var[T] $= 1/\lambda^2$
Quantil p (t_p)	$t_p = -\frac{1}{\lambda}\log(1-p)$

2. Distribuição Weibull

Este modelo de dois parâmetros ($\eta > 0$ e $\beta > 0$) é altamente flexível. Sua função de taxa de falha, $h(t)$, é monótona, podendo ser crescente ($\beta > 1$), decrescente ($\beta < 1$), ou constante ($\beta = 1$, reduzindo-se à Exponencial).

Tabela 2: Propriedades da Distribuição de Weibull

Propriedade	**Fórmula**
Função Densidade de Probabilidade ($f(t)$)	$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^{\beta}}$
Função de Sobrevivência ($S(t)$)	$S(t) = e^{-(t/\eta)^{\beta}}$
Função Taxa de Risco ($h(t)$)	$h(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}$
Média ($E[T]$)	$E[T] = \eta\Gamma\left(1+\frac{1}{\beta}\right)$
Variância (Var[T])	$\mathrm{Var}[T] = \eta^2\left[\Gamma\left(1+\frac{2}{\beta}\right) - \left(\Gamma\left(1+\frac{1}{\beta}\right)\right)^2\right]$
Quantil p (t_p)	$t_p = \eta[-\log(1-p)]^{1/\beta}$

*$\Gamma(\cdot)$ é a função Gamma.

3. Distribuição Log-Normal

A distribuição Log-Normal (LN) é caracterizada por dois parâmetros (μ e σ), sendo frequentemente utilizada quando o logaritmo natural do tempo de vida ($\log T$) segue uma distribuição normal. Sua taxa de falha é não monótona, aumentando inicialmente e depois diminuindo (forma unimodal).

Tabela 3: Propriedades da Distribuição Log-Normal

Propriedade	**Fórmula**
Função Densidade de Probabilidade ($f(t)$)	$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2}, \quad t > 0$
Função de Sobrevivência ($S(t)$)	$S(t) = 1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right)$
Função Taxa de Risco ($h(t)$)	$h(t) = \frac{f(t)}{1-\Phi\left(\frac{\log(t)-\mu}{\sigma}\right)}$
Média ($E[T]$)	$E[T] = e^{\mu+\sigma^2/2}$
Variância (Var[T])	$\mathrm{Var}[T] = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$
Quantil p (t_p)	$t_p = e^{\mu+z_p\sigma}$

*$\Phi(\cdot)$ é a FDA da normal padrão e $z_p = \Phi^{-1}(p)$.

6.8 Estimação dos Parâmetros dos Modelos

Cada modelo probabilístico é caracterizado por parâmetros desconhecidos (θ) que devem ser estimados a partir das observações amostrais. Devido à inapropriação do método dos Mínimos Quadrados para dados de tempo de vida (em especial pela dificuldade de incorporar a censura), o método de escolha é a Máxima Verossimilhança(Colosimo, 2024).

7 Método de Máxima Verossimilhança

Seja θ o vetor de parâmetros desconhecidos e $t_1, t_2, \ldots, t_n$ a amostra de observações. No contexto de sobrevivência, a presença de dados censurados implica que as observações contribuem para a verossimilhança através da função densidade (f) (para falhas) ou da função de sobrevivência (S) (para censuras).

Assumindo que em uma amostra de tamanho n, há r observações não censuradas e $n - r$ observações censuradas, a função de verossimilhança assume a seguinte forma:

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

Alternativamente, e de forma compacta, a função de verossimilhança pode ser definida com o uso da variável indicadora de censura δ_i (em que $\delta_i = 1$ para falha e $\delta_i = 0$ para censura):

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

A verossimilhança também pode ser reescrita utilizando a função taxa de falha $\lambda(t) = h(t)$:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [\lambda(t_i; \boldsymbol{\theta})]^{\delta_i} S(t_i; \boldsymbol{\theta})$$

O objetivo é encontrar o valor $\hat{\boldsymbol{\theta}}$ que maximize a função $L(\boldsymbol{\theta})$ (ou equivalentemente, a função log-verossimilhança, $\log L(\boldsymbol{\theta})$)(Colosimo, 2024).

7.1 Precisão das Estimativas e Intervalo de Confiança

Os estimadores de Máxima Verossimilhança ($\hat{\boldsymbol{\theta}}$) são assintoticamente normais. A precisão desses estimadores é avaliada pela sua variância assintótica.

A Matriz de Variância-Covariância dos estimadores de MV é aproximadamente o negativo da inversa da Matriz de Informação de Fisher, pois a variância do estimador de máxima verossimilhaça é geralmente desconhecida de forma fechada, $I(\boldsymbol{\theta})$:

$$\text{Var}(\hat{\boldsymbol{\theta}}) \approx [I(\boldsymbol{\theta})]^{-1} = \left[-E\left(\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right)\right]^{-1}$$

Quando o cálculo da Esperança é complexo, utilizamos a matriz Hessiana, avaliada nos estimadores:

$$\hat{\text{Var}}(\hat{\boldsymbol{\theta}}) = \left[-\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]^{-1}$$

Para a construção do Intervalo de Confiança (IC) de $100(1-\alpha)\%$ para um parâmetro genérico θ_i, utiliza-se a aproximação normal($\alpha = 0.05$):

$$IC(\theta_i) = \hat{\theta}_i \pm z_{\alpha/2} \times \sqrt{\hat{\text{Var}}(\hat{\theta}_i)}$$

onde $\sqrt{\hat{\text{Var}}(\hat{\theta}_i)}$ é o erro padrão estimado de $\hat{\theta}_i$.

7.2 Escolha do Modelo Probabilístico

A escolha do modelo probabilístico é uma etapa crucial na análise paramétrica de dados de tempo de vida, visto que o estimador de Máxima Verossimilhança (MV) só pode ser aplicado após a seleção de um modelo adequado. A decisão sobre qual modelo é o "melhor"para explicar os dados amostrais baseia-se na comparação entre valores estimados e observados. De acordo com a literatura, os métodos mais comuns são o método dos gráficos e o teste de hipóteses com modelos encaixados.

Método do Gráfico 1: Comparação da Função de Sobrevivência

Este método se baseia na comparação visual da função de sobrevivência estimada do modelo proposto, $\hat{S}_{\text{modelo}}(t)$, com o estimador não paramétrico de Kaplan-Meier, $\hat{S}_{\text{KM}}(t)$. O modelo considerado adequado será aquele cuja curva de sobrevivência $\hat{S}_{\text{modelo}}(t)$

se aproximar mais da curva de $\hat{S}_{\mathrm{KM}}(t)$ no mesmo gráfico.

Uma forma mais robusta deste método envolve a linearização da função taxa de falha acumulada, $\Lambda(t)$, a qual é relacionada à função de sobrevivência por:

$$\Lambda(t) = -\log(S(t))$$

Ao aplicar transformações específicas, é possível verificar se o modelo se ajusta aos dados pela linearidade dos pontos no gráfico(Colosimo, 2024).

Linearização para Distribuições Específicas

Distribuição Exponencial

Para o modelo exponencial, a transformação que lineariza a relação $\ln(S(t)$ e t é:

$$\ln S(t) = \ln\left(\exp(-\lambda t)\right) = -\lambda t.$$

Essa linearização que justifica o uso do gráfico semilog para verificar se o modelo exponencial é realmente adequado.

Distribuição Gumbel

No modelo gumbel a tranformação que lineariza a relação $\Lambda(t)$ é

$$\ln\left[-\ln S(t)\right] = \frac{1}{\sigma}\, t - \frac{\mu}{\sigma}.$$

Se os pontos se alinham aproximadamente dem uma reta, significa que há uma indicação de que a distribuiçãp gumbel é plausível para os tempos de sobrevida.

Distribuição de Weibull

Para o modelo de Weibull, a transformação que lineariza a relação $\Lambda(t)$ é a do logaritmo duplo:

$$\log[-\log(\hat{S}(t))] = \beta \cdot \log(t) - \beta \cdot \log(\eta)$$

Se o modelo for adequado, o gráfico de $\log[-\log(\hat{S}_{\mathrm{KM}}(t))]$ versus $\log(t)$ deve resultar em uma linha reta, onde β (parâmetro de forma) é o coeficiente angular(Aitkin e Clayton, 1980).

Distribuição Log-Normal

Para o modelo Log-Normal, a linearização é obtida através da transformação do escore normal:

$$\Phi^{-1}[1 - \hat{S}(t)] = \frac{1}{\sigma} \cdot \log(t) - \frac{\mu}{\sigma}$$

Em que Φ^{-1} é a função quantil da distribuição normal padrão. Se o modelo for adequado, o gráfico de $\Phi^{-1}[1 - \hat{S}_{\text{KM}}(t)]$ versus $\log(t)$ deve formar uma linha reta(Colosimo, 2024).

Método 2: Teste da Razão de Verossimilhança (TRV)

O Teste da Razão de Verossimilhança é uma técnica de teste de hipóteses que permite comparar formalmente a adequação de um modelo de interesse (H_0) versus um modelo alternativo mais geral ou encaixador (H_a). Este teste é aplicável apenas a modelos encaixados (ou aninhados).

Formulação e Estatística do Teste

As hipóteses são tipicamente definidas como:

- H_0: O modelo de interesse (mais simples/restrito) é adequado.
- H_a: O modelo encaixador (mais geral/complexo) é preferível.

A estatística do teste, TRV, é calculada com base nos valores máximos da função de log-verossimilhança obtidos pelo ajuste dos dois modelos:

$$TRV = -2 \log \left[\frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}}_G)} \right] = 2 \left[\log L(\hat{\boldsymbol{\theta}}_G) - \log L(\hat{\boldsymbol{\theta}}_0) \right]$$

Onde:

- $L(\hat{\boldsymbol{\theta}}_G)$ é a verossimilhança máxima do modelo Geral (encaixador).
- $L(\hat{\boldsymbol{\theta}}_0)$ é a verossimilhança máxima do modelo de Interesse (restrito).

Distribuição e Decisão

Sob a hipótese nula (H_0), a estatística TRV converge para uma distribuição Qui-Quadrado (χ^2). O número de graus de liberdade (GL) é dado pela diferença no número

de parâmetros entre os modelos

$$GL = (\text{Número de parâmetros em } \boldsymbol{\theta}_G) - (\text{Número de parâmetros em } \boldsymbol{\theta}_0).$$

Se o valor do TRV for superior ao valor crítico da distribuição χ^2 (ou se o p-valor for menor que o nível de significância α), rejeita-se H_0. Isto indica que a simplificação do modelo (modelo de interesse) não é justificada e o modelo mais geral deve ser escolhido. Caso contrário, não havendo evidências para rejeitar H_0, o modelo mais simples(Colosimo, 2024).

8 Estimação por Máxima Verossimilhança

Nos últimos vinte anos, aproximadamente, ocorreram avanços significativos na qualidade do ajuste de modelos de misturas finitas por meio do método da Máxima Verossimilhança (MV). Até então, havia uma grande dificuldade em ajustar modelos de mistura a dados multidimensionais, principalmente devido à presença de múltiplos máximos locais na função de verossimilhança que é um dos temas que será abordado mais adiante neste trabalho.

A estimação por máxima verossimilhança funciona da seguinte forma: sejam os dados observados

$$Y_1, Y_2, \ldots, Y_n.$$

Escolhe-se uma função de probabilidade (ou densidade) que se acredita representar adequadamente o comportamento desses dados, de acordo com o modelo assumido. Em seguida, busca-se o valor do vetor de parâmetros Ψ que maximiza a função de verossimilhança, ou seja, os parâmetros que mais "explicam" os dados observados.

Após estimar os parâmetros, é necessário avaliar o quão precisas são essas estimativas. Para isso, utilizam-se as matrizes de informação de Fisher.

8.1 Matriz de Informação de Fisher

A matriz de informação de Fisher mede a quantidade de informação que os dados fornecem sobre os parâmetros do modelo. Ela é definida como

$$I(\Psi) = \mathbb{E}_\Psi \left[S(Y;\Psi) S^T(Y;\Psi) \right],$$

em que

$$S(Y;\Psi) = \frac{\partial \log L(\Psi)}{\partial \Psi}$$

é o vetor escore, isto é, o gradiente da log-verossimilhança em relação aos parâmetros.

A matriz de informação de Fisher também pode ser interpretada como o negativo da esperança da Hessiana da log-verossimilhança, sendo, portanto, uma medida da curvatura dessa função. Intuitivamente:

- quando a verossimilhança é muito "pontuda" (alta curvatura), os parâmetros são estimados com maior precisão;
- quando é mais "achatada", há maior incerteza associada às estimativas.

8.2 Matriz de Covariância Assintótica

Uma vez obtido o estimador de máxima verossimilhança $\hat{\Psi}$, é necessário quantificar sua variabilidade. Sob condições regulares e para grandes amostras, a matriz de covariância assintótica do estimador é dada por

$$\mathrm{Cov}(\hat{\Psi}) \approx [I(\Psi)]^{-1}.$$

Na prática, como a matriz de informação esperada $I(\Psi)$ nem sempre é de fácil obtenção, utiliza-se a matriz de informação observada, calculada diretamente a partir dos dados:

$$I(\hat{\Psi}; y) = -\frac{\partial^2 \log L(\hat{\Psi})}{\partial \Psi \, \partial \Psi^T}.$$

Assim, o erro padrão de cada parâmetro $\hat{\Psi}_r$ pode ser aproximado pela raiz quadrada do elemento correspondente na diagonal principal da inversa da matriz de informação observada:

$$SE(\hat{\Psi}_r) \approx \sqrt{\left[I^{-1}(\hat{\Psi}; y)\right]_{rr}}.$$

Portanto, a matriz de informação desempenha papel fundamental tanto na inferência quanto na avaliação da precisão das estimativas obtidas por máxima verossimilhança(Colosimo, 2024).

9 O Algoritmo Expectation-Maximization

Algoritmo Expectation-Maximization (EM) é um método iterativo e robusto em estatística, fundamentalmente desenvolvido para encontrar estimativas de Máxima Verossimilhança (MV) em modelos estatísticos que contêm variáveis latentes (não observadas) ou dados incompletos(Silva, 2013).

9.1 Relação Fundamental de Probabilidade

A relação que conecta a densidade dos dados completos com os observados e latentes.

$$f(z|\psi) = f(x, y|\psi) = k(y|x, \psi)f(x|\psi).$$

9.2 Log-Verossimilhança em Termos dos Dados Completos

A log-verossimilhança dos dados observados, $l(\psi|x)$, expressa através de uma decomposição logarítmica.

$$l(\psi|x) = \log[f(X, Y|\psi)] - \log[k(Y|X, \psi)].$$

9.3 Aplicação da Expectativa

A equação da log-verossimilhança após aplicar a Expectativa sobre os dados ausentes.

$$l(\psi|x) = E_{Y|X,\psi^{(t)}}[\log f(X, Y|\psi)] - E_{Y|X,\psi^{(t)}}[\log k(Y|X, \psi)].$$

Onde o subscrito $Y|X, \psi^{(t)}$ indica que a expectativa é tomada sobre a distribuição condicional de Y, dados X e os parâmetros atuais $\psi^{(t)}$.

9.4 Definição da Função Q

A definição da Função Q, que é a parte central do passo E (Expectation) e o objetivo da maximização no passo M.

$$Q(\psi|\psi^{(t)}) = E[\log f(X, Y|\psi)|X, \psi^{(t)}].$$

9.5 Definição da Função H

A definição da Função H, que corresponde à segunda parte da decomposição.

$$H(\psi|\psi^{(t)}) = E[\log k(Y|X,\psi)|X,\psi^{(t)}].$$

Decomposição Final da Log-Verossimilhança

Juntando tudo, a log-verossimilhança pode ser decomposta em termos das funções Q e H.

$$l(\psi|x) = Q(\psi|\psi^{(t)}) - H(\psi|\psi^{(t)}).$$

A necessidade do EM surge porque, na presença desses dados "ocultos", a função de verossimilhança direta se torna analiticamente intratável ou muito complexa para ser maximizada. O EM resolve isso operando em um ciclo repetitivo de dois passos que garantem a convergência para um pico local (máximo) da verossimilhança:

- **Passo E (Expectation – Expectativa):** Com base nos parâmetros atuais do modelo, o algoritmo calcula a expectativa da verossimilhança completa (a verossimilhança que teríamos se todos os dados estivessem visíveis). Essencialmente, é feita uma "melhor suposição"probabilística sobre os dados ou variáveis latentes que faltam.
- **Passo M (Maximization – Maximização):** O algoritmo, então, maximiza a verossimilhança esperada calculada no Passo E, como se ela fosse a verossimilhança real dos dados completos. Isso gera um novo e melhor conjunto de estimativas de parâmetros para o modelo.

Para entender por que a maximização iterativa de $Q(\Psi;\Psi^{(s-1)})$ é suficiente para garantir um aumento em $l(\Psi|\mathcal{X})$, analisamos a diferença na log-verossimilhança observada entre duas iterações sucessivas, $\Psi^{(s)}$ e $\Psi^{(s-1)}$.Temos

$$\begin{aligned} l(\Psi^{(s)}|\mathcal{X}) - l(\Psi^{(s-1)}|\mathcal{X}) = &[Q(\Psi^{(s)};\Psi^{(s-1)}) - Q(\Psi^{(s-1)};\Psi^{(s-1)})] \\ &+ [H(\Psi^{(s)};\Psi^{(s-1)}) - H(\Psi^{(s-1)};\Psi^{(s-1)})] \end{aligned}$$

O segundo termo entre colchetes, $[H(\Psi^{(s)};\Psi^{(s-1)}) - H(\Psi^{(s-1)};\Psi^{(s-1)})]$, corres-

ponde ao negativo da divergência de Kullback-Leibler (KL)(Oliveira (2017). Pela desigualdade de Jensen, demonstra-se que esta diferença é sempre não-positiva:

$$\begin{aligned} H(\Psi, \Psi^{(s-1)}) - H(\Psi^{(s-1)}, \Psi^{(s-1)}) &= E\left[\log\left(\frac{f(\mathcal{Y}|\mathcal{X}, \Psi)}{f(\mathcal{Y}|\mathcal{X}, \Psi^{(s-1)})}\right)\Big|\mathcal{X}, \Psi^{(s-1)}\right] \\ &\leq \log\left(E\left[\frac{f(\mathcal{Y}|\mathcal{X}, \Psi)}{f(\mathcal{Y}|\mathcal{X}, \Psi^{(s-1)})}\Big|\mathcal{X}, \Psi^{(s-1)}\right]\right) \\ &= \log\left(\int f(\mathcal{Y}|\mathcal{X}, \Psi)d\mathcal{Y}\right) \leq 0. \end{aligned}$$

(Nota: O valor final da integral é 1, resultando em $\log(1) = 0$*, o que comprova que a desigualdade de Jensen garante que a diferença é* ≤ 0*.)*

concluímos que o segundo termo é ≤ 0. Além disso, o Passo M garante que $\Psi^{(s)}$ é escolhido tal que $Q(\Psi^{(s)}; \Psi^{(s-1)}) \geq Q(\Psi^{(s-1)}; \Psi^{(s-1)})$. Combinando estas informações, obtemos a propriedade de convergência do EM

$$l(\Psi^{(s)}|\mathcal{X}) - l(\Psi^{(s-1)}|\mathcal{X}) \geq Q(\Psi^{(s)}; \Psi^{(s-1)}) - Q(\Psi^{(s-1)}; \Psi^{(s-1)}) \geq 0.$$

Assim, () demonstra que qualquer escolha de $\Psi^{(s)}$ que melhore a função Q em relação a $\Psi^{(s-1)}$ terá um efeito igual ou maior no aumento da log-verossimilhança observada $l(\Psi|\mathcal{X})$.

Isto garante que a sequência de estimadores $\{\Psi^{(s)}\}$ converge para um ponto crítico (máximo local ou saddle point) de $l(\Psi|\mathcal{X})$. Concluímos que a diferença $l(\Psi^{(s)}|\mathcal{X}) - l(\Psi^{(s-1)}|\mathcal{X})$ nunca é negativa.

O ciclo se repete ($E \to M \to E \to M...$) até que as estimativas dos parâmetros convirjam. Este processo é amplamente utilizado em diversas áreas da estatística e aprendizado de máquina, sendo a aplicação mais notável nos Modelos de Mistura de Distribuições, onde a variável latente é a qual subpopulação cada observação pertence(Silva, 2013).

Modelagem de Regressão para Tempos de Falha: O Modelo AFT

Para a modelagem de tempos de falha sujeitos a censura, frequentemente utiliza-se a suposição de Tempo de Falha Acelerado (AFT). Nesse modelo, o logaritmo do tempo de falha é vinculado linearmente a um conjunto de covariáveis acrescido de um termo de erro aleatório(Kleinbaum, D. G.; Klein, 2012):

$$\log T_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \sigma \epsilon_i.$$

O modelo AFT pode ser classificado como paramétrico ou semiparamétrico. No modelo paramétrico, assume-se uma distribuição específica para o erro ϵ_i. Já no modelo semiparamétrico, a distribuição do erro não é especificada, conferindo maior robustez e flexibilidade, embora seja mais complexo estimar certos parâmetros(SEO; KANG, 2021).

9.6 Estimativa do Intercepto e Robustez

Enquanto a análise de regressão geralmente foca nos coeficientes de inclinação ($\boldsymbol{\beta}$), a estimativa do parâmetro de intercepto (β_0) é fundamental para prever o tempo médio até a falha.

No modelo semiparamétrico, a consistência do estimador do intercepto depende de condições restritivas, como o suporte irrestrito para algumas covariáveis (Susarla e Van Ryzin, 1980; Ding e Nan, 2015). Por outro lado, no modelo paramétrico, a estimativa do intercepto não requer procedimentos separados ou condições adicionais. Apesar de o modelo paramétrico ser mais sensível a *outliers* e a desvios da distribuição assumida, ele foi o mais indicado para este estudo devido o caso parametrico ter à necessidade de uma grande amplitude de dados para a estimação do intercepto(SEO; KANG, 2021).

9.7 Inferência por Máxima Verossimilhança

No modelo paramétrico, a distribuição imposta sobre ϵ_i define a distribuição de T_i. Distribuições típicas para o erro incluem a de valor extremo, normal e logística, que resultam nas distribuições de **Weibull, log-normal e log-logística** para o tempo de falha, respectivamente.

Para dados com censura à direita aleatória, a função de verossimilhança é construída como(SEO; KANG, 2021).

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[f(\log y_i; \boldsymbol{\theta})\right]^{\delta_i} \left[S(\log y_i; \boldsymbol{\theta})\right]^{1-\delta_i}.$$

Onde $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma)^\top$, $f(\cdot)$ é a função de densidade e $S(\cdot)$ é a função de sobrevivência. O Estimador de Máxima Verossimilhança (MLE), $\hat{\boldsymbol{\theta}}$, é obtido maximizando esta função. Sob condições de regularidade, $\hat{\boldsymbol{\theta}}$ apresenta distribuição assintoticamente normal:

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1}).$$

Onde $I(\boldsymbol{\theta})^{-1}$ é a inversa da matriz de informação de Fisher. O MLE fornece diretamente $\hat{\beta}_0$, permitindo estimar o valor previsto do logaritmo do tempo para um dado conjunto de covariáveis $\mathbf{x}$:

$$\hat{E}[\log T|\mathbf{x}] = \hat{\beta}_0 + \mathbf{x}^\top \hat{\boldsymbol{\beta}}^*.$$

10 Modelo de Mistura AFT

No contexto da análise de sobrevivência, a heterogeneidade da população pode ser modelada assumindo que os indivíduos pertencem a diferentes subgrupos latentes. O modelo de mistura(ZÖRNIG, 2024) atribui diferentes distribuições de latência para cada grupo, permitindo que indivíduos reajam de formas distintas ao tratamento ou permaneçam livres do evento por longos períodos.

Seja Y uma variável latente indicadora da subpopulação, tal que a probabilidade de um indivíduo pertencer à j-ésima subpopulação é dada por $P(Y = j) = \pi_j$. A função de densidade de probabilidade (f.d.p.) marginal do tempo de falha t, condicionada a um vetor de covariáveis $\mathbf{x}$, é expressa como

$$f(t|\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{K} \pi_j(\mathbf{x}, \boldsymbol{\rho}) \cdot f_j(t|\mathbf{x}, \boldsymbol{\phi}_j).$$

Em que:

- $\boldsymbol{\theta} = (\boldsymbol{\rho}, \boldsymbol{\phi})$ representa o conjunto de todos os parâmetros desconhecidos do modelo;

- $\pi_j(\mathbf{x}, \boldsymbol{\rho})$ é a probabilidade de incidência para a j-ésima subpopulação;
- $f_j(t|\mathbf{x}, \boldsymbol{\phi}_j)$ é a densidade de latência condicional do tempo de falha para o grupo j, assumida contínua e indexada pelo vetor de parâmetros $\boldsymbol{\phi}_j$.

A função de sobrevivência correspondente para este modelo de mistura é dada por(KUO; PENG, 1995)

$$S(t|\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{K} \pi_j(\mathbf{x}, \boldsymbol{\rho}) \cdot S_j(t|\mathbf{x}, \boldsymbol{\phi}_j).$$

10.1 Incidência via Regressão Logística

Para um modelo com K componentes, a probabilidade de incidência π_j para os grupos é modelada pela função *softmax*, garantindo que $\sum_{j=1}^{K} \pi_j = 1$,(KUO; PENG, 1995).

$$\pi_j(\mathbf{x}, \boldsymbol{\rho}) = \frac{\exp(\mathbf{x}'\boldsymbol{\rho}_j)}{\sum_{l=1}^{K} \exp(\mathbf{x}'\boldsymbol{\rho}_l)}.$$

Nesta formulação:

- $\pi_j(\mathbf{x}, \boldsymbol{\rho})$: é a probabilidade de um indivíduo com características $\mathbf{x}$ pertencer ao subgrupo j.
- $\mathbf{x}$: é o vetor de covariáveis.
- $\boldsymbol{\rho}_j$: são os coeficientes de regressão que determinam como as covariáveis influenciam a probabilidade de alocação naquele grupo específico.
- $\boldsymbol{\rho}_K = \mathbf{0}$: por convenção de identificabilidade, o último grupo é utilizado como referência.

10.2 Função de Log-Verossimilhança

Considerando uma amostra de n observações independentes (t_i, δ_i), onde δ_i é o indicador de falha (1 para evento observado e 0 para censura), a função de log-verossimilhança para a estimação dos parâmetros é definida como(KUO; PENG, 1995)

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{K} \pi_j(\mathbf{x}_i, \boldsymbol{\rho}) \left[f_j(t_i|\mathbf{x}_i, \boldsymbol{\phi}_j)^{\delta_i} \cdot S_j(t_i|\mathbf{x}_i, \boldsymbol{\phi}_j)^{1-\delta_i} \right] \right).$$

10.3 Verossimilhança Completa

A verossimilhança completa, condicional às variáveis latentes, é (Dempster, Laird & Rubin 1977)

$$L_c(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\sigma}) = \prod_{i=1}^{n} \prod_{j=1}^{J} \left[\pi_j(\boldsymbol{z}_i, \boldsymbol{\gamma})^{\eta_{ij}} \cdot \left(\frac{1}{\sigma_j t_i} f_\epsilon(u_{ij}) \right)^{\delta_i \eta_{ij}} \cdot (S_\epsilon(u_{ij}))^{(1-\delta_i)\eta_{ij}} \right].$$

11 Estimação via Algoritmo EM

O algoritmo EM (Expectation-Maximization) é utilizado para encontrar os estimadores de máxima verossimilhança na presença de dados faltantes. Neste contexto, as variáveis indicadoras de grupo Z_{ij} (se o indivíduo i pertence ao grupo j) são consideradas latentes (Dempster, Laird & Rubin 1977).

11.1 Verossimilhança Completa

Seja η_{ij} uma variável indicadora latente tal que $\eta_{ij} = 1$ se o i-ésimo indivíduo pertence ao grupo j, e 0 caso contrário. A função de verossimilhança completa é dada por:

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{K} \left[\pi_j(\mathbf{x}_i, \boldsymbol{\rho}) \cdot \left(\frac{1}{\sigma_j t_i} f_\epsilon(u_{ij}) \right)^{\delta_i} (S_\epsilon(u_{ij}))^{1-\delta_i} \right]^{\eta_{ij}}$$

Onde $u_{ij} = \frac{\log t_i - \mathbf{x}_i' \boldsymbol{\beta}_j}{\sigma_j}$ é o resíduo padronizado, f_ϵ é a densidade e S_ϵ é a sobrevivência da distribuição do erro.

11.2 Passo E (Expectation)

Nesta etapa, calcula-se a esperança condicional das variáveis latentes η_{ij}, dado os dados observados e as estimativas correntes dos parâmetros $\boldsymbol{\theta}^{(k)}$ na iteração k. Essa esperança corresponde à probabilidade posterior de pertinência ao grupo j, denotada por $w_{ij}^{(k)}$ (Wei, G. C. G., & Tanner, M. A., 1990):

$$w_{ij}^{(k)} = E[\eta_{ij} \mid \text{dados}, \boldsymbol{\theta}^{(k)}] = \frac{\pi_j^{(k)} \cdot \left(\frac{1}{\sigma_j^{(k)} t_i} f_\epsilon(u_{ij}^{(k)})\right)^{\delta_i} \left(S_\epsilon(u_{ij}^{(k)})\right)^{1-\delta_i}}{\sum_{l=1}^{K} \pi_l^{(k)} \cdot \left(\frac{1}{\sigma_l^{(k)} t_i} f_\epsilon(u_{il}^{(k)})\right)^{\delta_i} \left(S_\epsilon(u_{il}^{(k)})\right)^{1-\delta_i}}.$$

11.3 Passo M (Maximization)

Maximiza-se a esperança da log-verossimilhança completa ponderada pelos pesos $w_{ij}^{(k)}$ obtidos no passo anterior. A maximização pode ser separada em duas partes independentes.

Incidência

Para os parâmetros da mistura, maximiza-se a função Q_1

$$Q_1(\boldsymbol{\rho}) = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij}^{(k)} \log \pi_j(\mathbf{x}_i, \boldsymbol{\rho}).$$

Esta etapa equivale a ajustar uma regressão logística multinomial ponderada, em que os pesos são dados por $w_{ij}^{(k)}$.

Latência

Para cada componente $j = 1, \ldots, K$, maximiza-se a função Q_2 referente aos parâmetros de tempo de vida

$$Q_2(\boldsymbol{\beta}_j, \sigma_j) = \sum_{i=1}^{n} w_{ij}^{(k)} \left[\delta_i \log\left(\frac{1}{\sigma_j} f_\epsilon(u_{ij})\right) + (1-\delta_i) \log S_\epsilon(u_{ij})\right].$$

$$Q_2(\boldsymbol{\beta}_j, \sigma_j) \propto \sum_{i=1}^{n} w_{ij}^{(k)} \left[\delta_i \log f_\epsilon(u_{ij}) - \delta_i \log \sigma_j + (1 - \delta_i) \log S_\epsilon(u_{ij})\right].$$

A maximização desta função corresponde à estimação de um modelo AFT ponderado para a subpopulação j. Para distribuições como Weibull ou Log-Normal, métodos numéricos como Newton-Raphson são frequentemente utilizados(AITKIN, M.; CLAYTON, D., 1980).

12 Metodologia

O estudo utiliza a base de dados públicos de Chicco (2020), composta por 299 pacientes com insuficiência cardíaca e fração de ejeção reduzida. As variáveis de resposta para a análise de sobrevivência são `time` (tempo de acompanhamento em dias) e `DEATH_EVENT` (indicador de falha/censura, δ). A principal covariável explanatória investigada é a `ejection_fraction` (%). O pré-processamento dos dados foi realizado para garantir a adequação aos modelos.

12.1 Fluxo de Modelagem Estatística

A análise foi conduzida em quatro etapas principais, utilizando a linguagem R e Python em etapas diferentes e bibliotecas como `lifelines`, `auton-survival` e `scipy.optimize`.

Inicialmente, estimou-se a curva de sobrevivência global da amostra usando o estimador de Kaplan-Meier (KM). Em seguida, a amostra foi estratificada pelos grupos clínicos de FEVE (Reduzida, Intermediária, Preservada) para avaliar o poder prognóstico da classificação tradicional e identificar visualmente a heterogeneidade.A validade do modelo de Cox foi condicionada à verificação do pressuposto de Riscos Proporcionais (RP). A satisfação desta premissa, que exige que o efeito das covariáveis seja constante ao longo do tempo, foi avaliada pelos Resíduos de Schoenfeld. A análise visual dos gráficos e o teste estatístico global foram empregados para determinar a robustez do modelo.

Em caso de violação significativa do pressuposto de RP por alguma covariável,seria considerada a aplicação de um Modelo AFT (Accelerated Failure Time) de Misturas. Este modelo paramétrico alternativo lida diretamente com o tempo de falha, sendo mais adequado para situações onde os riscos não são proporcionais e quando a amostra é composta por subgrupos distintos que não são completamente representados pela suposição de homogeneidade de um único modelo de Cox.

Etapa 2: Modelagem Paramétrica (Componente Único, $K = 1$)

Para testar a adequação de distribuições paramétricas, foram ajustados modelos Acelerados de Tempo de Falha (AFT) com componente único ($K = 1$), assumindo homogeneidade da amostra. As distribuições Exponencial, Weibull, Lognormal e Gumbel foram testadas. A seleção da melhor distribuição-base foi feita por comparação gráfica

com a curva KM e pelo menor valor do Critério de Informação de Akaike (AIC). O modelo Lognormal apresentou o melhor ajuste.

Etapa 3: Modelagem de Mistura (Heterogeneidade, $K > 1$)

Para capturar a heterogeneidade bimodal observada na FEVE, foram ajustados Modelos de Mistura AFT (AFT-MM) com $K = 2$ e $K = 3$, componentes da mesma familía, utilizando as distribuições Weibull, Gumbel e Lognormal. A estimação dos parâmetros (pesos π_k e parâmetros da distribuição) foi realizada por Máxima Verossimilhança via algoritmo Expectation-Maximization (EM).

Etapa 4: Seleção e Validação do Modelo Final

A seleção do modelo final foi baseada em dois critérios:

1. **Ajuste Estatístico:** O Critério de Informação Bayesiano (BIC) foi usado para selecionar o modelo mais parcimonioso (melhor equilíbrio entre ajuste e complexidade).
2. **Qualidade da Separação:** A análise visual da responsabilidade (probabilidade de pertencimento ao cluster) que fornece a separação mais "limpa"e menos ambígua dos pacientes em subgrupos.

Finalmente, a validade clínica verificada. Os pacientes foram alocados aos seus respectivos clusters. Curvas de Kaplan-Meier foram geradas para cada cluster e o teste Log-Rank foi aplicado para quantificar a diferença na sobrevida entre eles.

12.2 Avaliação e Vantagem Metodológica

A comparação entre os modelos semiparamétricos, paramétricos e de mistura AFT será realizada com base em medidas de ajuste e desempenho preditivo. A qualidade do ajuste será avaliada por meio dos critérios **AIC** e **BIC**(Paula,2013). Além do desempenho quantitativo, destaca-se a vantagem metodológica dos modelos de mistura AFT em termos de interpretação e identificação de heterogeneidade clínica. Esses modelos permitem revelar subgrupos latentes de pacientes com padrões de sobrevida distintos, proporcionando uma visão mais detalhada e personalizada dos riscos associados à IC. As curvas de sobrevivência específicas por subgrupo evidenciarão a diversidade de comporta-

mentos clínicos e a capacidade dos modelos de estimar riscos diferenciados, frequentemente mascarados por abordagens tradicionais. Assim, a metodologia proposta combina rigor estatístico,consistência clínica e potencial interpretativo elevado, consolidando-se como uma alternativa robusta para a análise de sobrevida em pacientes com insuficiência cardíaca.

12.3 Conjunto de dados

A base de dados utilizada neste estudo foi extraída do trabalho de chicco2020, intitulado "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone", publicado na revista *BMC Medical Informatics and Decision Making*. O conjunto de dados contém informações de 299 pacientes com insuficiência cardíaca, coletadas no Instituto de Cardiologia de Faisalabad e no Hospital Allied, localizados em Faisalabad (Punjab, Paquistão), durante o período de abril a dezembro de 2015. Todos os pacientes apresentavam (FEVE). A amostra é composta por 105 mulheres e 194 homens, todos com histórico de insuficiência cardíaca.

A base de dados é formada por 13 variáveis que descrevem características clínicas e de estilo de vida dos pacientes, incluindo medidas laboratoriais e indicadores de saúde relevantes para a análise da sobrevida.

12.4 Descrição das variáveis do conjunto de dados

A base de dados é composta por 13 variáveis que descrevem informações demográficas, clínicas e laboratoriais dos pacientes. A Tabela 4 apresenta o nome, o tipo e a descrição de cada variável.

Tabela 4: Descrição das variáveis do conjunto de dados de pacientes com insuficiência cardíaca.

Variável	Tipo	Descrição
`age`	Numérica (anos)	Idade do paciente em anos completos.
`anaemia`	Binária (0 = Não, 1 = Sim)	Presença de anemia, definida pela redução do nível de hemoglobina no sangue.
`creatinine_phosphokinase`	Numérica (mcg/L)	Nível de creatina fosfoquinase (CPK) medido no sangue, indicador de dano muscular, incluindo lesão cardíaca.
`diabetes`	Binária (0 = Não, 1 = Sim)	Indica se o paciente possui diagnóstico de diabetes mellitus.
`ejection_fraction`	Numérica (%)	Fração de ejeção do ventrículo esquerdo, medida em porcentagem, que representa a eficiência de bombeamento do coração.
`high_blood_pressure`	Binária (0 = Não, 1 = Sim)	Indica se o paciente tem histórico de hipertensão arterial sistêmica.
`platelets`	Numérica (kiloplatelets/mL)	Contagem de plaquetas no sangue, medida em milhares por mililitro.
`serum_creatinine`	Numérica (mg/dL)	Concentração de creatinina sérica, indicador da função renal.
`serum_sodium`	Numérica (mEq/L)	Concentração de sódio no soro, associada ao equilíbrio eletrolítico.
`sex`	Binária (0 = Feminino, 1 = Masculino)	Sexo biológico do paciente.
`smoking`	Binária (0 = Não, 1 = Sim)	Indica se o paciente é fumante.
`time`	Numérica (dias)	Tempo de acompanhamento do paciente, em dias, até o óbito ou censura.
`DEATH_EVENT`	Binária (0 = Censura, 1 = Falha)	Variável de desfecho que indica se o paciente faleceu durante o período de acompanhamento.

Fonte: Adaptado de chicco2020.

Tabela 5: Primeiras seis observações do conjunto de dados analisado.

Idade	Anemia	Fosfoquinase	Diabetes	Fração de ejeção	Pressão alta
75	0	582	0	20	1
55	0	7861	0	38	0
65	0	146	0	20	0
50	1	111	0	20	0
65	1	160	1	20	0
90	1	47	0	40	1

Fonte: Elaboração própria com base nos dados da pesquisa feito no R.

Tabela 6: Estatísticas Descritivas das Variáveis Contínuas

Variável	Média	Mediana	Desvio Padrão	Variância	Mínimo (p0)	Máximo (p100)
Idade	60.83	60	11.89	141.37	40	95
Fosfoquinase de creatinina	581.84	250	970.29	941463.79	23	7861
Fração de ejeção	38.08	38	11.83	140.06	14	80
Plaquetas	263358.03	262000	97804.24	9565662700	25100	850000
Creatinina sérica	1.39	1.1	1.03	1.06	0.5	9.4
Sódio sérico	136.63	137	4.41	19.45	113	148
Tempo _dias	130.26	115	77.61	6023.31	4	285

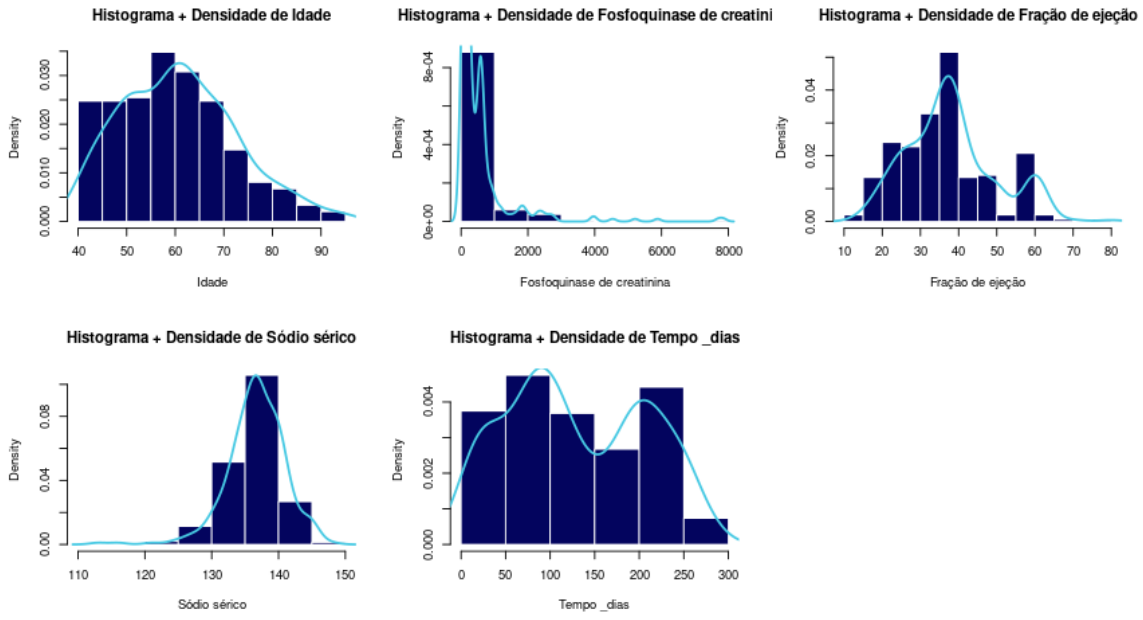
Fonte: Elaboração própria com base nos dados da pesquisa feito no R.

Tabela 7: Resumo das Variáveis Fatoriais (Categóricas)

Variável	n_missing	complete_rate	ordered	n_unique	top_counts
Anemia	0	1	FALSE	2	Não: 170, Sim: 129
diabetes	0	1	FALSE	2	Não: 174, Sim: 125
Pressão alta	0	1	FALSE	2	Não: 194, Sim: 105
Sexo	0	1	FALSE	2	Mas: 194, Fem: 105
Fumando	0	1	FALSE	2	Não: 203, Sim: 96
Evento_de_morte	0	1	FALSE	2	Censura: 203, Falha: 96

Fonte: Elaboração própria com base nos dados da pesquisa feito no R.

13 Aplicação do estudo



13.1 Análise Descritiva das Variáveis Contínuas

Distribuição de Idade: A distribuição de idade é razoavelmente simétrica e unimodal, e é um pouco parecida com a distribuição normal (Gaussiana). Conseguimos verificar que a maioria dos pacientes da amostra se concentra na faixa dos 50 aos 70 anos, com pacientes com 60 anos que mais se sobressaem. Isso indica que o estudo é composto predominantemente por pacientes idosos.

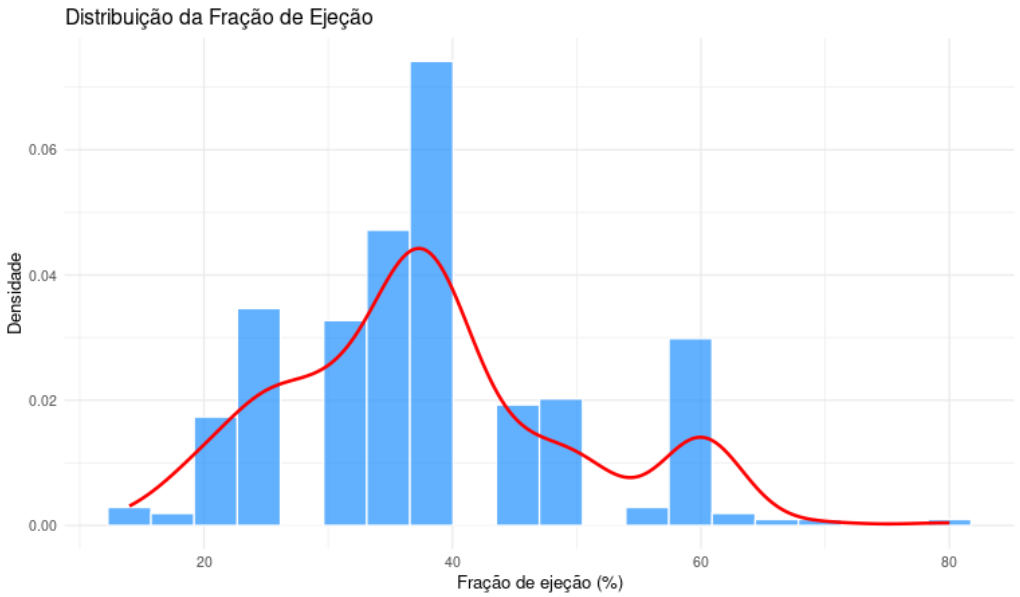
Distribuição de Fração de Ejeção: A distribuição de fração de ejeção é claramente bimodal, ou seja, tem dois picos, o que foi muito de encontro com o estudo. Isso indica que a amostra não é homogênea, ou seja, há dois subgrupos claros de pacientes.

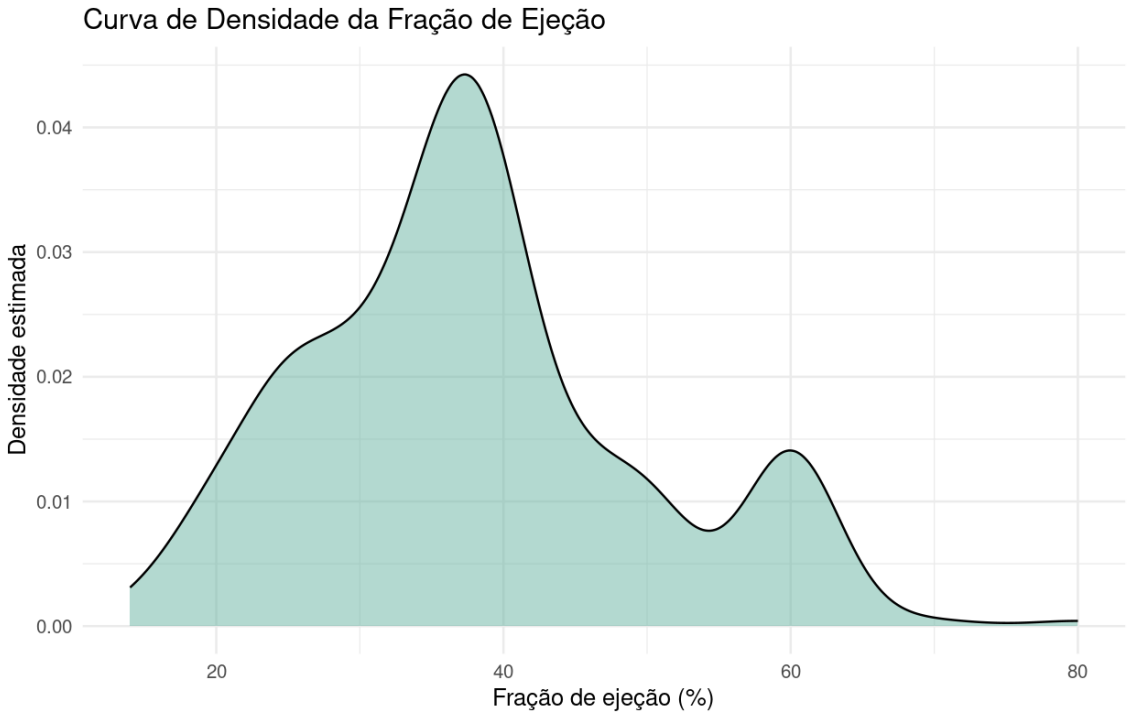
Distribuição de Creatinina Sérica: A distribuição de Creatinina Sérica é fortemente assimétrica à direita. A grande maioria dos pacientes tem níveis baixos de creatinina, no entanto, existe uma "cauda longa"de pacientes com valores muito altos (outliers).

Distribuição de Sódio Sérico: A distribuição de sódio sérico é a que mais se aproxima de uma distribuição normal é unimodal. O nível de sódio sérico dos pacientes é bem comportado e homogêneo, com a média centrada em aproximadamente 137 mEq/L.

Distribuição do tempo de dias: Já a distribuição do tempo de dias tem uma grande variabilidade entre os pacientes, possui valores distribuidos aproximadamente entre 0 e 270 dias, há uma leve assimetria à direita, refletindo a presença de alguns tempos mais longos, embora em menor quantidade.

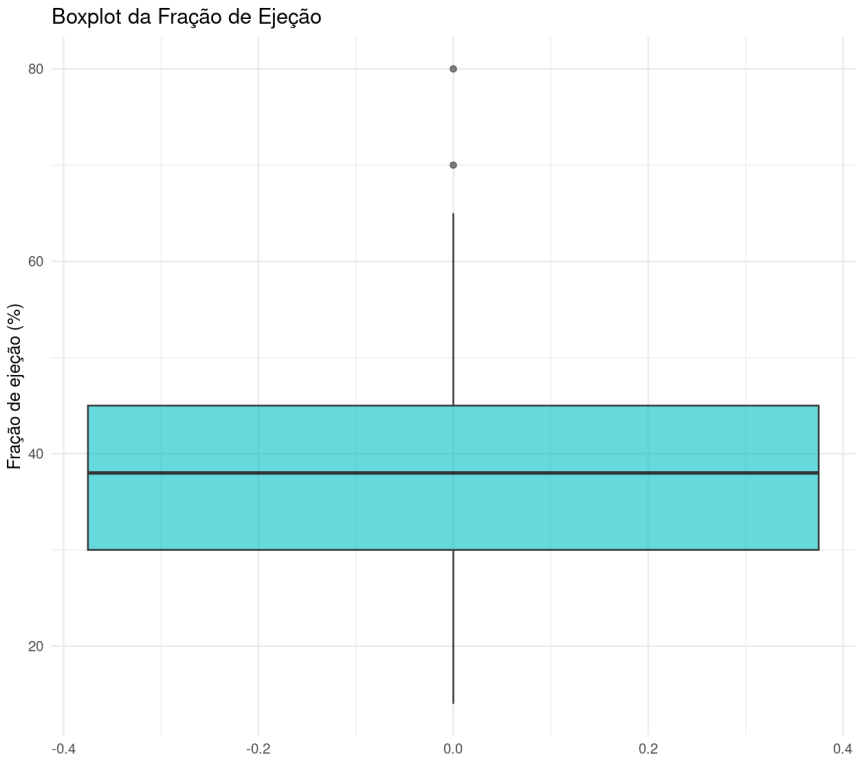
13.2 Análise da variável principal do estudo - FRAÇÃO DE EJEÇÃO





13.3 Justificativa da Bimodalidade

Como mostrado nos gráficos de densidade e histograma das variáveis contínuas, a distribuição da Fração de Ejeção (FEVE) revela uma característica crucial: ambos os gráficos mostram inequivocamente que a distribuição não é unimodal, mas sim bimodal. Esta bimodalidade, com um pico em torno de 38% e outro em 60%, confirma que a amostra de pacientes é estatisticamente heterogênea.



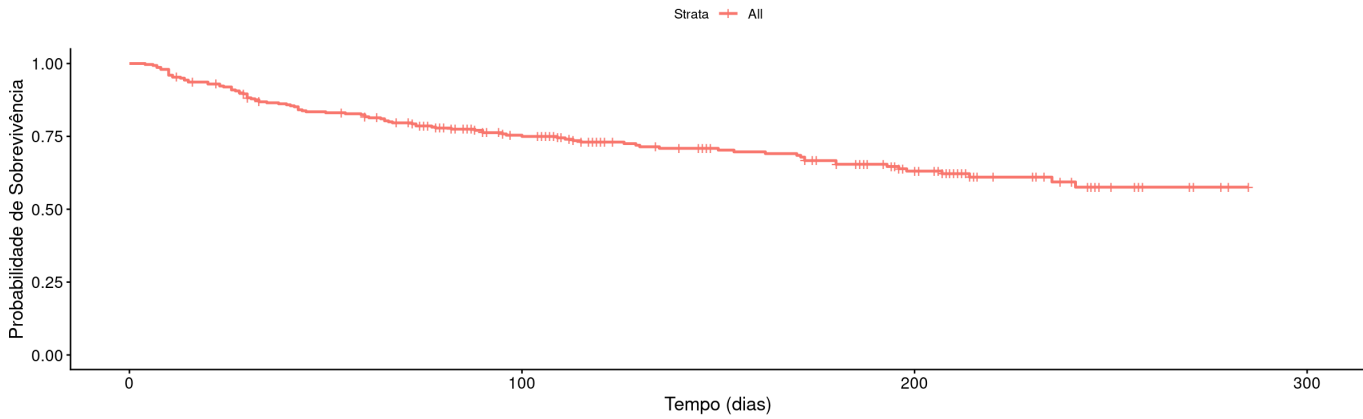
13.4 Análise do Boxplot da Fração de Ejeção

O boxplot da Fração de Ejeção fornece um resumo estatístico claro. A mediana (a linha central) está localizada ligeiramente abaixo de 40%, aproximadamente em 38%. A caixa, que representa a dispersão central (o Intervalo Interquartil ou IQR), indica que 50% de todos os pacientes possuem uma fração de ejeção entre aproximadamente 30% (Q1) e 45% (Q3). Além disso, o gráfico identifica claramente dois valores atípicos (outliers), representados pelos pontos individuais, em níveis muito altos: um em aproximadamente 70% e outro em 80%.

13.5 Análise Não Paramétrica de Sobrevivência

Na etapa inicial da análise exploratória dos dados, ao examinar as variáveis quantitativas, foi observado um comportamento bimodal na variável fração de ejeção (FE), que constitui a principal variável de interesse deste estudo. A presença de dois picos na função de densidade pode auxiliar na constução dos grupos latentes.

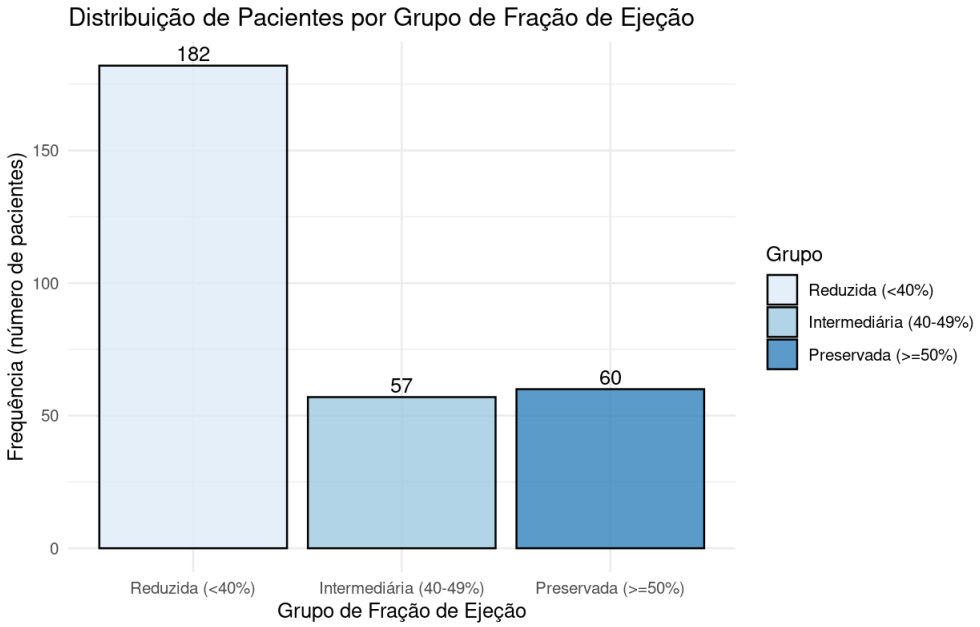
Inicialmente, foi ajustado um modelo de sobrevivência semiparamétrico, baseado no estimador de Kaplan-Meier mostrado abaixo:



A curva de sobrevivência geral mostrou uma redução constante ao longo do tempo, com uma queda acentuada nos primeiros 50 dias. Observou-se ainda que o tempo mediano de sobrevivência não foi atingido durante o período de acompanhamento, indicando uma censura considerável dos dados. O gráfico de Kaplan-Meier teve papel fundamental como ponto de partida para a análise paramétrica posterior.

Em seguida, os pacientes foram classificados conforme a fração de ejeção ventricular esquerda (FEVE) em três grupos:





13.6 Descrição do Boxplot da Fração de Ejeção (FE)

O gráfico apresentado é um Boxplot (Diagrama de Caixa) que compara a distribuição da Fração de Ejeção (FE) em porcentagem (eixo y) em três grupos distintos definidos pela própria faixa de FE (eixo x): Reduzida ($< 40\%$), Intermediária ($40 - 50\%$) e Preservada ($\geq 50\%$).

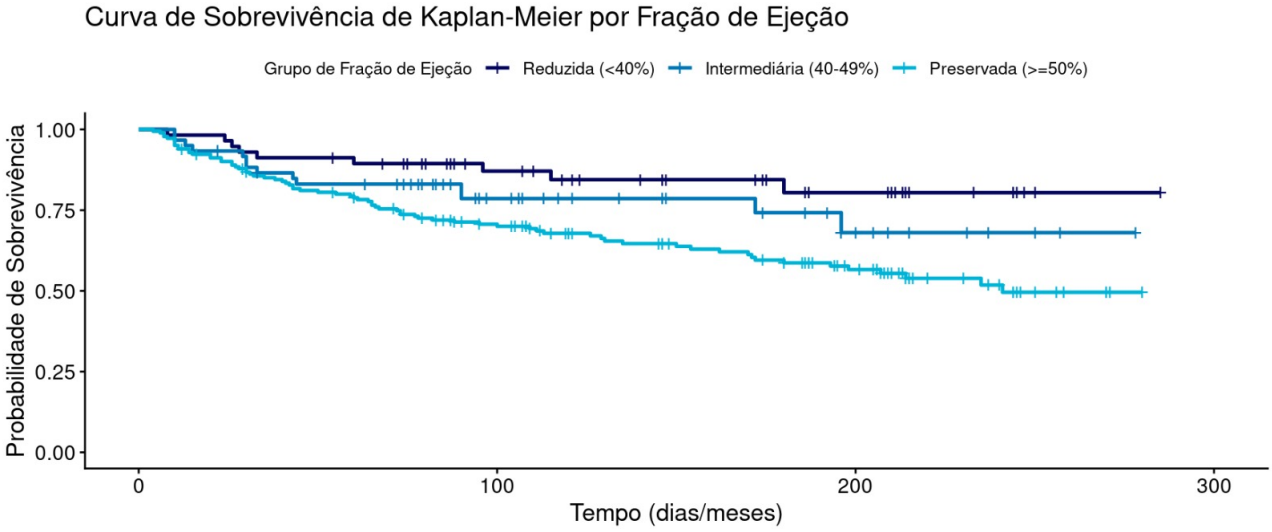
A análise por grupo revela as seguintes características:

- **Grupo Reduzida ($< 40\%$)**: Este grupo, representado pela caixa vermelha, exibe os valores mais baixos de FE. A mediana (linha central) está aproximadamente em

30%. O Intervalo Interquartil (IIQ), que abrange 50% dos dados centrais, se estende de cerca de 25% a 35%.

- **Grupo Intermediária** $(40 - 49\%)$: Caracterizado pela caixa verde, este grupo apresenta a menor dispersão, uma vez que está restrito à faixa de definição. A **mediana** se localiza no meio deste intervalo, perto de **45**%. Notavelmente, toda a distribuição parece estar contida dentro dos limites da caixa, sem bigodes visíveis, indicando que a grande maioria dos dados se encontra exatamente entre 40% e 49%.

- **Grupo Preservada** $(\geq 50\%)$: Este grupo, em azul, concentra os valores mais elevados de FE. A **mediana** se situa em torno de **55**%. O IIQ varia aproximadamente de 50% a 60%. O grupo apresenta a maior dispersão geral e inclui um **valor atípico (outlier)** distinto, representado por um ponto isolado acima de 60%, indicando um valor de FE excepcionalmente alto.



Kaplan-Meier foi aplicada nos três grupos distintos também, conforme descrito a seguir:

- **Reduzida** $(< 40\%)$ – **Azul Escuro:** grupo majoritário $(n = 182)$, apresentou a melhor probabilidade de sobrevivência, mantendo-se próxima de 80% após 285 dias.

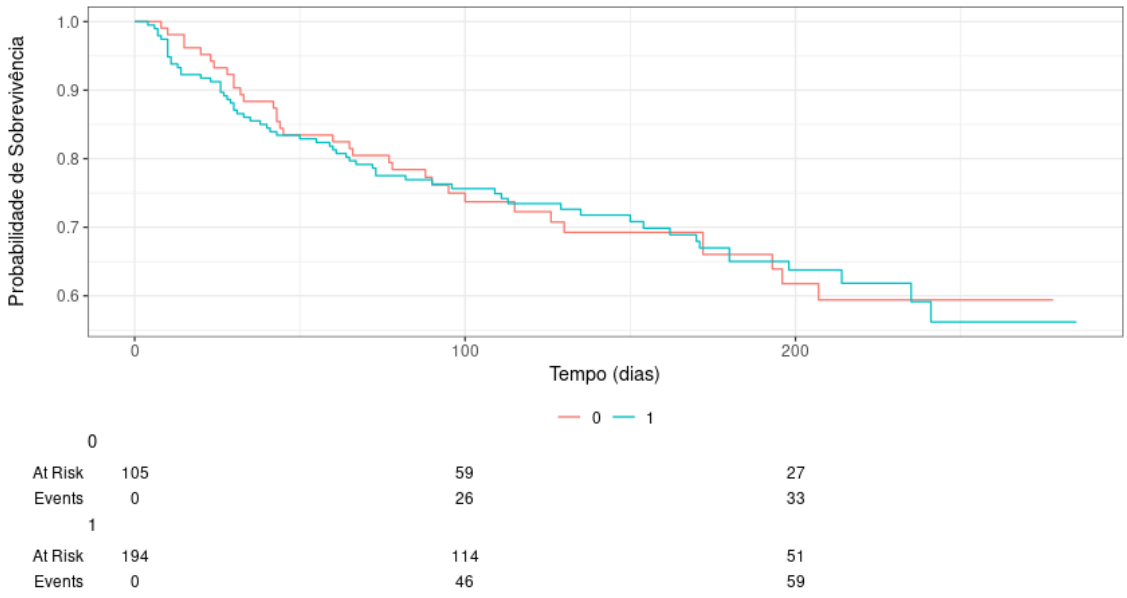
- **Intermediária** $(40–49\%)$ – **Azul Médio:** grupo com prognóstico intermediário $(n = 57)$, encerrando o período com cerca de 70% de sobrevivência.

- **Preservada ($\geq$ 50%) – Ciano/Azul Claro:** grupo ($n = 60$) com o pior prognóstico, sendo o único cuja curva cruzou a linha de 50% de sobrevivência, indicando que metade dos pacientes evoluiu para óbito antes do término do acompanhamento.

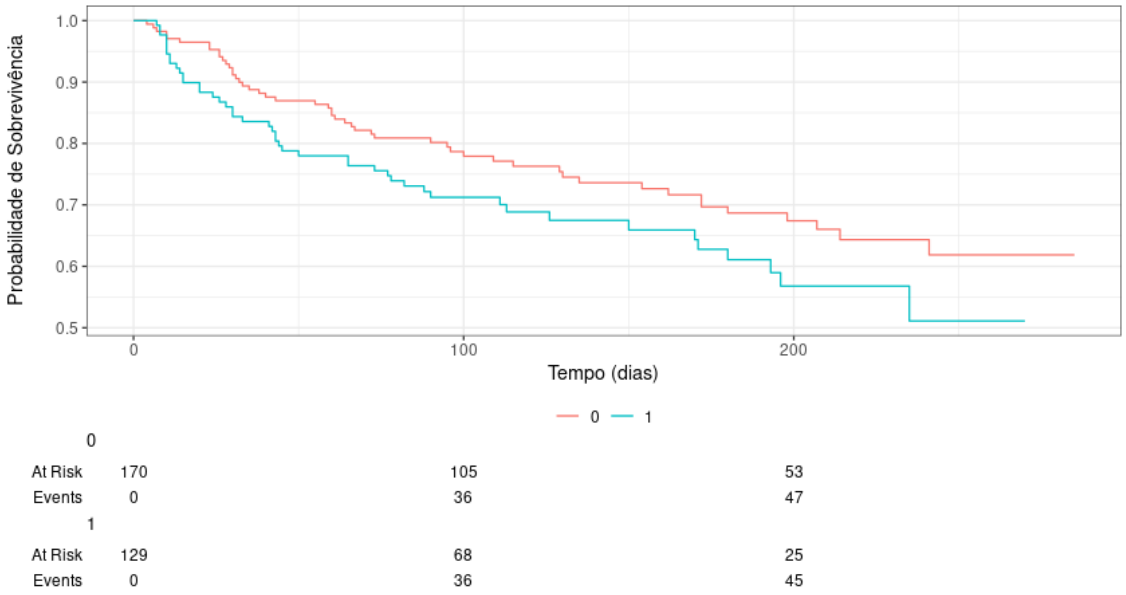
Esse resultado é contraintuitivo e paradoxal, uma vez que pacientes com FE preservada ($\geq$ 50%) geralmente apresentam melhor função cardíaca. Tal comportamento reforça a hipótese de heterogeneidade não observada e sustenta a motivação para aplicação dos modelos de mistura.

Depois da análise de Kaplan-Meier, foi elaborado a modelagem de Preditores de Risco (Regressão de cox) para quantificar a influência simultânea e independente de múltiplas variáveis sobre o risco de mortalidade.Começando pela análise gráfica de Kaplan-Meier para variáveis categóricas para verificar se as curvas são proporcionais (Não se cruzam).

. Variável Sexo

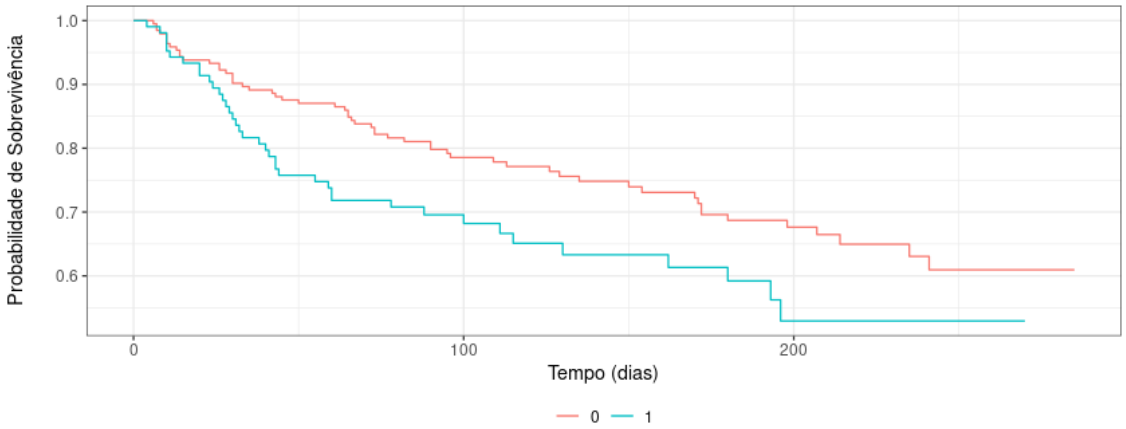


. Variável Anemia



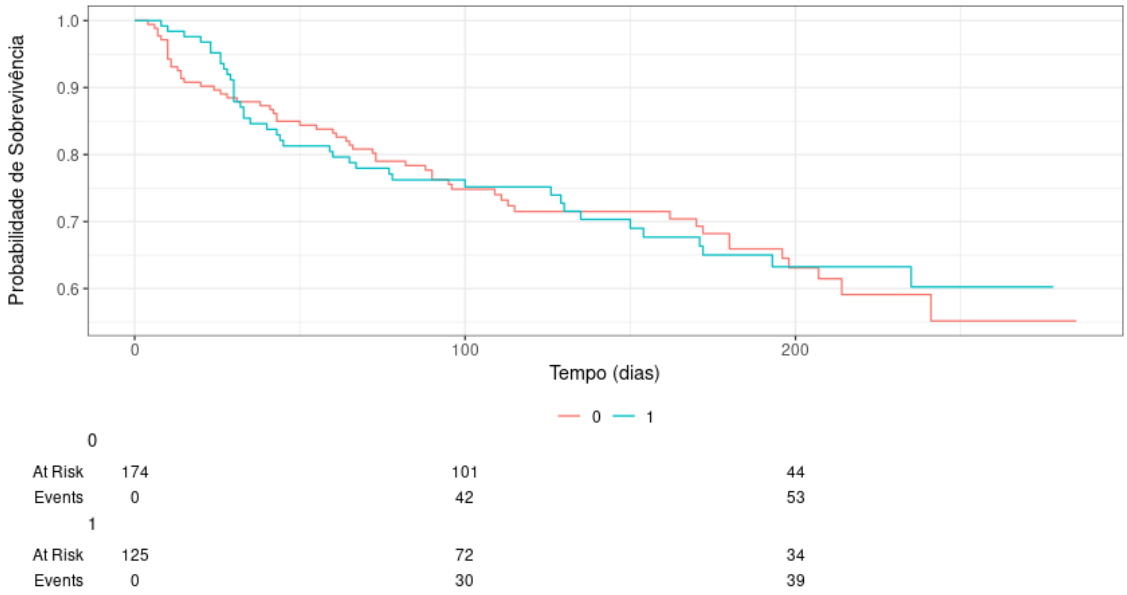
0			
At Risk	170	105	53
Events	0	36	47
1			
At Risk	129	68	25
Events	0	36	45

. Variável Pressão Alta

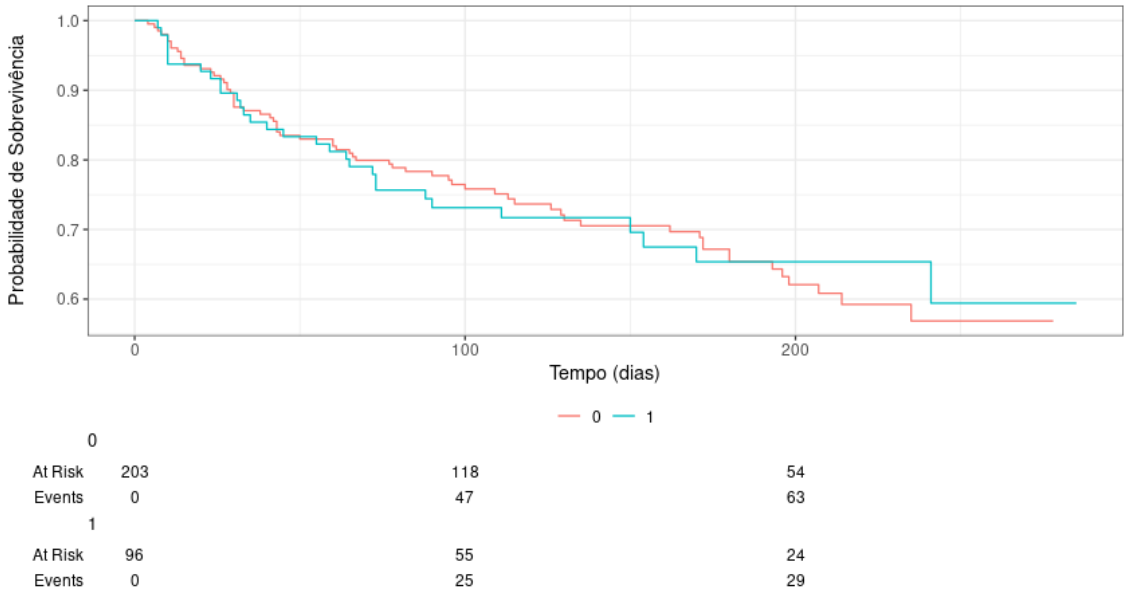


0			
At Risk	194	122	64
Events	0	40	53
1			
At Risk	105	51	14
Events	0	32	39

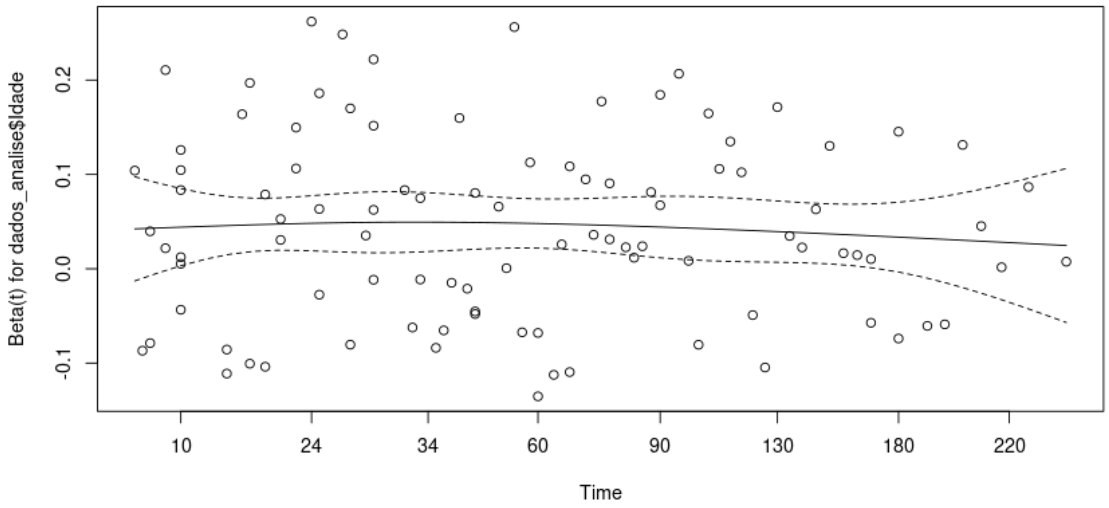
. Variável Diabetes

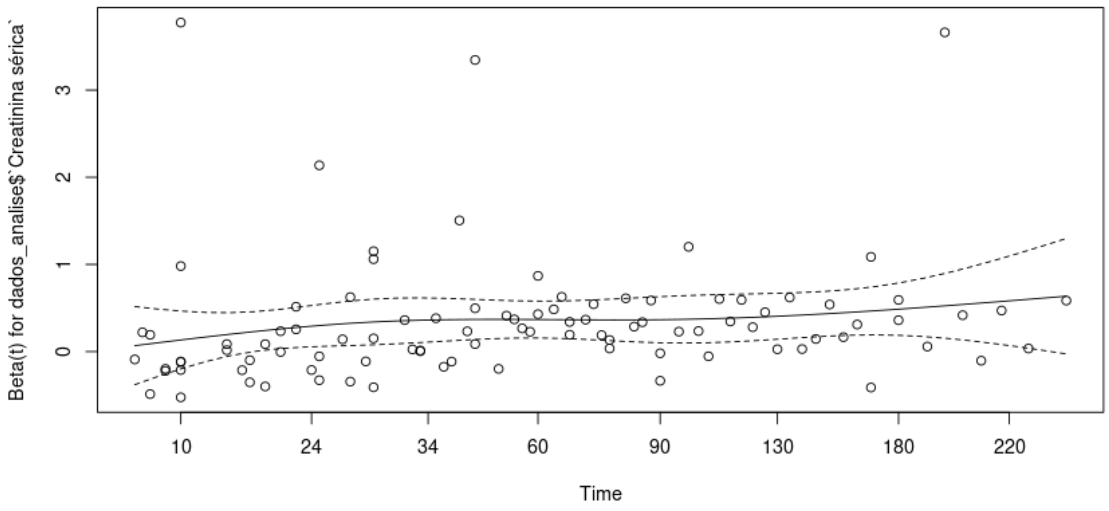


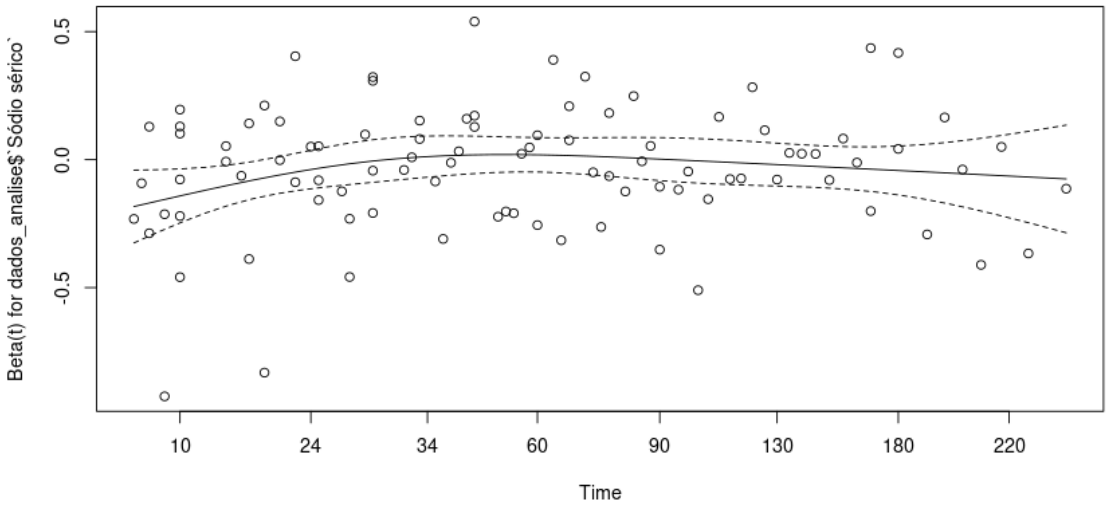
. Variável Fumante

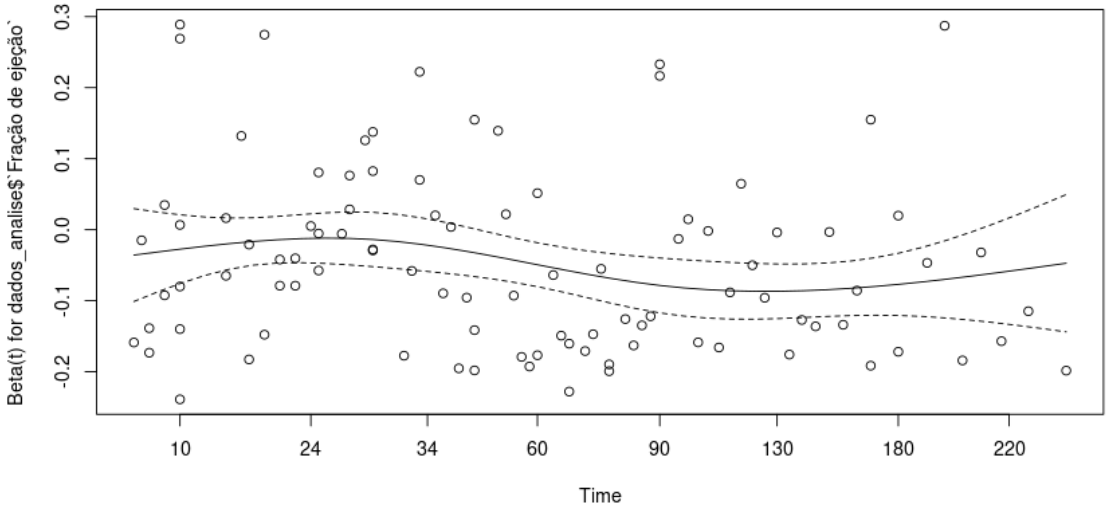


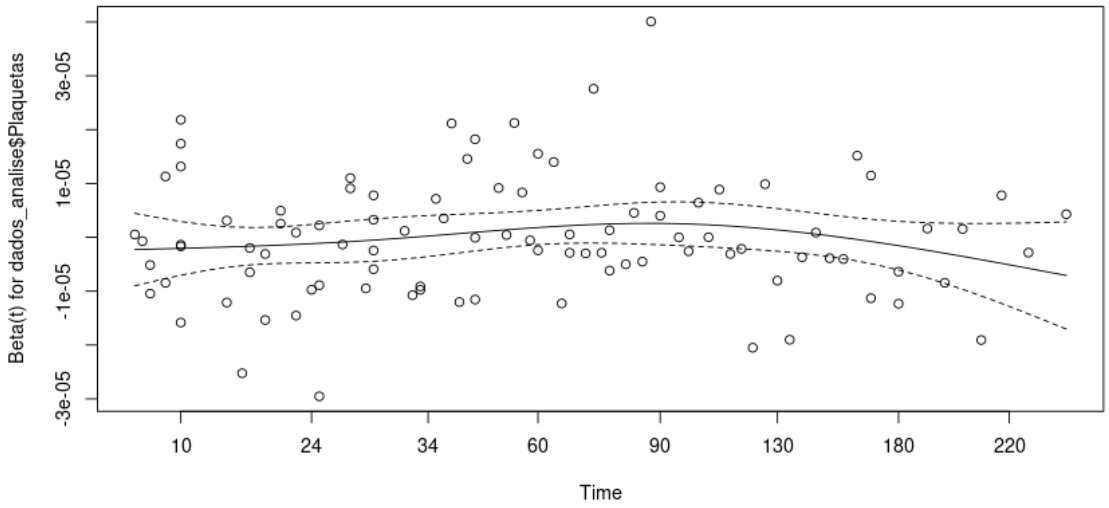
A avaliação do pressuposto de riscos proporcionais revelou comportamentos distintos entre as variáveis. Para as variáveis categóricas — especificamente Sexo, Diabetes e Tabagismo —, a inspeção visual das curvas de sobrevivência indicou violações do princípio de proporcionalidade, evidenciadas pelo cruzamento das curvas ao longo do tempo. Já para as variáveis contínuas, a validade do modelo foi verificada através da análise dos resíduos de Schoenfeld, assegurando que os efeitos estimados das covariáveis permanecessem constantes durante o período de acompanhamento.

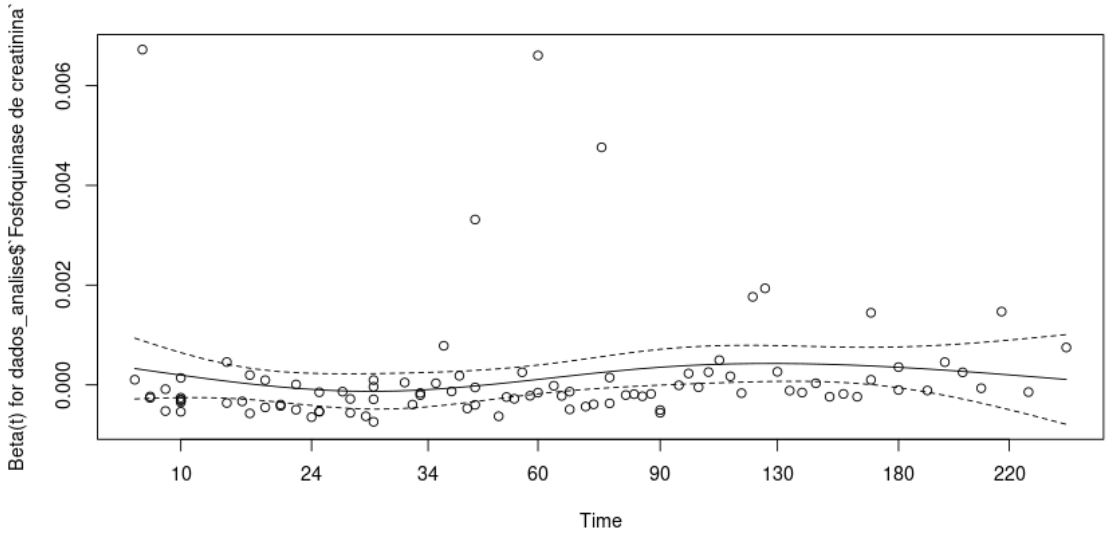

Beta(t) for dados_analise$Idade
0.2
0.1
0.0
-0.1
10
24
34
60
90
130
180
220
Time


Beta(t) for dados_analise$`Creatinina sérica`
3
2
1
0
10
24
34
60
90
130
180
220
Time


Beta(t) for dados_analises$`Sódio sérico`
0.5
0.0
-0.5
10
24
34
60
90
130
180
220
Time


Beta(t) for dados_analises$`Fração de ejeção`
0.3
0.2
0.1
0.0
-0.1
-0.2
10
24
34
60
90
130
180
220
Time





Visualmente, a Fosfoquinase de Creatinina, Creatinina Sérica e Plaquetas demonstraram o melhor ajuste à premissa de RP. Em ambos os casos, a linha de tendência (suavizada) dos resíduos se manteve consideravelmente linear e horizontal, e majoritariamente concentrada dentro dos limites de seus respectivos intervalos de confiança. Este comportamento indica que o efeito dessas variáveis no risco de mortalidade é constante ao longo do tempo, fortalecendo a inclusão delas no Modelo de Cox.

13.7 Criação do modelo de Cox

Nessa etapa foi feito um modelo de cox completo com todas as variáveis inclusas no modelo

```
  n= 299, number of events= 96

                                          coef  exp(coef)   se(coef)      z Pr(>|z|)
dados_analise$Idade                   4.655e-02  1.048e+00  9.320e-03  4.994 5.90e-07 ***
dados_analise$`Fosfoquinase de creatinina`  2.214e-04  1.000e+00  9.928e-05  2.230   0.0258 *
dados_analise$Plaquetas              -6.337e-07  1.000e+00  1.119e-06 -0.566   0.5712
dados_analise$`Creatinina sérica`     3.203e-01  1.378e+00  7.020e-02  4.562 5.06e-06 ***
dados_analise$`Sódio sérico`         -4.395e-02  9.570e-01  2.328e-02 -1.888   0.0591 .
dados_analise$`Fração de ejeção`     -4.900e-02  9.522e-01  1.047e-02 -4.678 2.89e-06 ***
dados_analise$Anemia                  4.618e-01  1.587e+00  2.169e-01  2.129   0.0332 *
dados_analise$diabetes                1.417e-01  1.152e+00  2.229e-01  0.636   0.5250
dados_analise$`Pressão alta`          4.789e-01  1.614e+00  2.163e-01  2.214   0.0268 *
dados_analise$Sexo                   -2.443e-01  7.832e-01  2.516e-01 -0.971   0.3315
dados_analise$Fumando                 1.330e-01  1.142e+00  2.511e-01  0.530   0.5964

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                      exp(coef) exp(-coef) lower .95 upper .95
dados_analise$Idade                      1.0476     0.9545    1.0287    1.0670
dados_analise$`Fosfoquinase de creatinina`    1.0002     0.9998    1.0000    1.0004
dados_analise$Plaquetas                  1.0000     1.0000    1.0000    1.0000
dados_analise$`Creatinina sérica`        1.3775     0.7259    1.2005    1.5807
dados_analise$`Sódio sérico`             0.9570     1.0449    0.9143    1.0017
dados_analise$`Fração de ejeção`         0.9522     1.0502    0.9328    0.9719
dados_analise$Anemia                     1.5869     0.6302    1.0374    2.4274
dados_analise$diabetes                   1.1523     0.8679    0.7444    1.7837
dados_analise$`Pressão alta`             1.6142     0.6195    1.0564    2.4666
dados_analise$Sexo                       0.7832     1.2768    0.4783    1.2825
dados_analise$Fumando                    1.1422     0.8755    0.6982    1.8685

Concordance= 0.742  (se = 0.027 )
Likelihood ratio test= 82.1  on 11 df,   p=6e-13
Wald test            = 87.28  on 11 df,   p=6e-14
Score (logrank) test = 88.26  on 11 df,   p=4e-14
```

A avaliação global do modelo demonstrou um ajuste robusto e estatisticamente significativo, conforme indicado pelos testes de significância global . O poder discriminatório do modelo foi considerado bom, apresentando um Índice de Concordância (C-index) de 0.742, sugerindo que o modelo é capaz de discriminar corretamente, em 74,2% das vezes, qual indivíduo em um par aleatório experimentará o evento de interesse (mortalidade) primeiro.

A análise evidenciou que a Idade e a Creatinina Sérica são os fatores de risco mais potentes para a mortalidade, ambos com alta significância estatística. Para cada ano adicional de idade, o risco de morte aumenta em aproximadamente 4,8% (HR = 1.048).

De maneira similar, o aumento de uma unidade na Creatinina Sérica eleva o risco de óbito em 37,8% (HR = 1.378), sublinhando a disfunção renal como um prognóstico crítico.

Em contrapartida, a **Fração de Ejeção** foi o principal fator de proteção identificado. Com um HR de 0.952, o risco de mortalidade é reduzido em cerca de 4,8% a cada aumento de um ponto percentual na Fração de Ejeção do Ventrículo Esquerdo.

Outras variáveis atingiram o limiar de significância estatística ($\alpha = 0.05$): a presença de Pressão Alta e Anemia aumentaram o risco em 61,4% e 58,7%, respectivamente. A Creatinina Fosfoquinase também foi significativa, mas com um HR muito próximo da unidade (HR = 1.0002), indicando relevância estatística, porém um impacto clínico marginal.

Variáveis como Diabetes, Tabagismo, Sexo e Plaquetas não demonstraram associação estatística independente com a mortalidade no contexto deste modelo multivariado. O Sódio Sérico apresentou um p-valor limítrofe ($p = 0.0591$), sugerindo uma tendência que não atingiu a significância convencional de 5%. Após a análise do Modelo de Cox completo, procedeu-se ao refinamento do modelo por meio da remoção sequencial das covariáveis que não demonstraram associação estatística independente e significativa com o desfecho de mortalidade (p-valor $> 0,05$). As variáveis *Sódio Sérico*, *Plaquetas*, *Diabetes*, *Sexo* e *Tabagismo* foram removidas do modelo final, resultando em um modelo mais parsimonioso.

```
                                    exp(coef) exp(-coef) lower .95 upper .95
dados_analise$Idade                    1.0445     0.9574    1.0266     1.063
dados_analise$`Fosfoquinase de creatinina`    1.0002     0.9998    1.0000     1.000
dados_analise$`Creatinina sérica`      1.3686     0.7306    1.1956     1.567
dados_analise$`Sódio sérico`           0.9553     1.0468    0.9126     1.000
dados_analise$`Fração de ejeção`       0.9536     1.0486    0.9346     0.973
dados_analise$Anemia                   1.5630     0.6398    1.0255     2.382
dados_analise$`Pressão alta`           1.6431     0.6086    1.0808     2.498

Concordance= 0.738  (se = 0.027 )
Likelihood ratio test= 80.58  on 7 df,   p=1e-14
Wald test            = 88.44  on 7 df,   p=3e-16
Score (logrank) test = 87.66  on 7 df,   p=4e-16
```

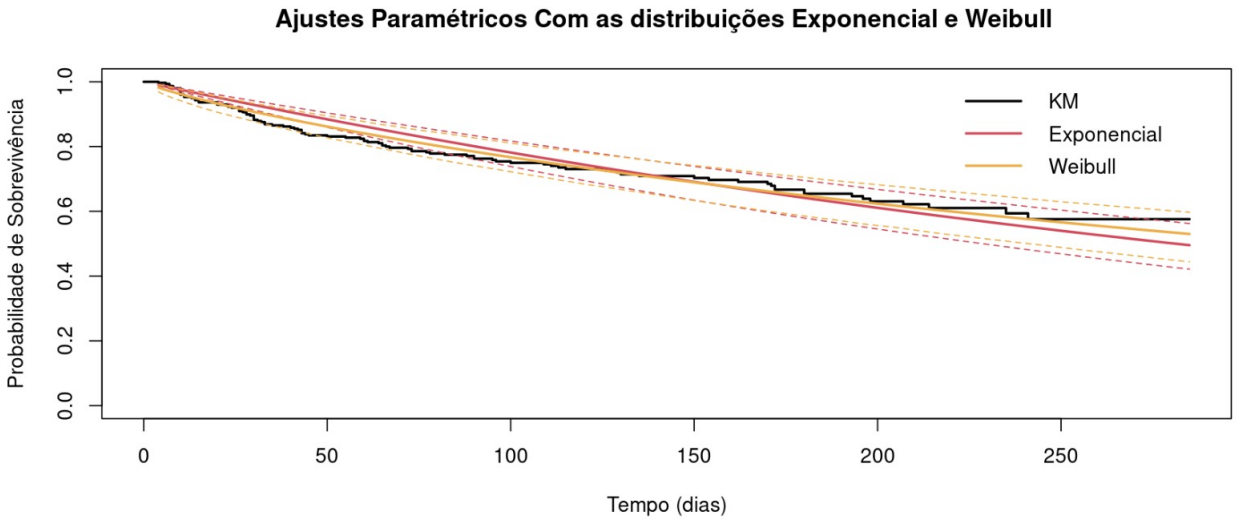
```
  n= 299, number of events= 96

                                          coef  exp(coef)   se(coef)      z Pr(>|z|)
dados_analise$Idade                   4.356e-02  1.045e+00  8.829e-03  4.934 8.07e-07 ***
dados_analise$`Fosfoquinase de creatinina`  2.100e-04  1.000e+00  9.826e-05  2.137   0.0326 *
dados_analise$`Creatinina sérica`     3.138e-01  1.369e+00  6.895e-02  4.551 5.33e-06 ***
dados_analise$`Sódio sérico`         -4.570e-02  9.553e-01  2.336e-02 -1.957   0.0504 .
dados_analise$`Fração de ejeção`     -4.747e-02  9.536e-01  1.027e-02 -4.621 3.82e-06 ***
dados_analise$Anemia                  4.466e-01  1.563e+00  2.150e-01  2.077   0.0378 *
dados_analise$`Pressão alta`          4.966e-01  1.643e+00  2.137e-01  2.324   0.0201 *
---
```

O modelo refinado manteve as variáveis que demonstraram contribuição estatisticamente relevante (p-valor $\leq 0,05$) e foi reajustado, consolidando os preditores independentes de risco, conforme apresentado.

13.8 Análise Paramétrica de Sobrevivência

Posteriormente, foram ajustados modelos de sobrevivência paramétricos, com base nas distribuições Exponencial, Weibull, Lognormal e Gumbel (Imagens abaixo). Os gráficos compararam as curvas de sobrevivência observadas (Kaplan-Meier) com as curvas ajustadas por cada modelo. Verificou-se que os modelos Lognormal e Gumbel apresentaram o melhor ajuste visual, aproximando-se da trajetória empírica da curva de Kaplan-Meier. O modelo Exponencial mostrou-se o menos adequado, incapaz de capturar a curvatura real da função de sobrevivência, enquanto o modelo Weibull demonstrou boa flexibilidade e aderência aos dados. O modelo Gumbel apresentou ajuste satisfatório, embora tenha subestimado ligeiramente a sobrevivência entre 100 e 200 dias.



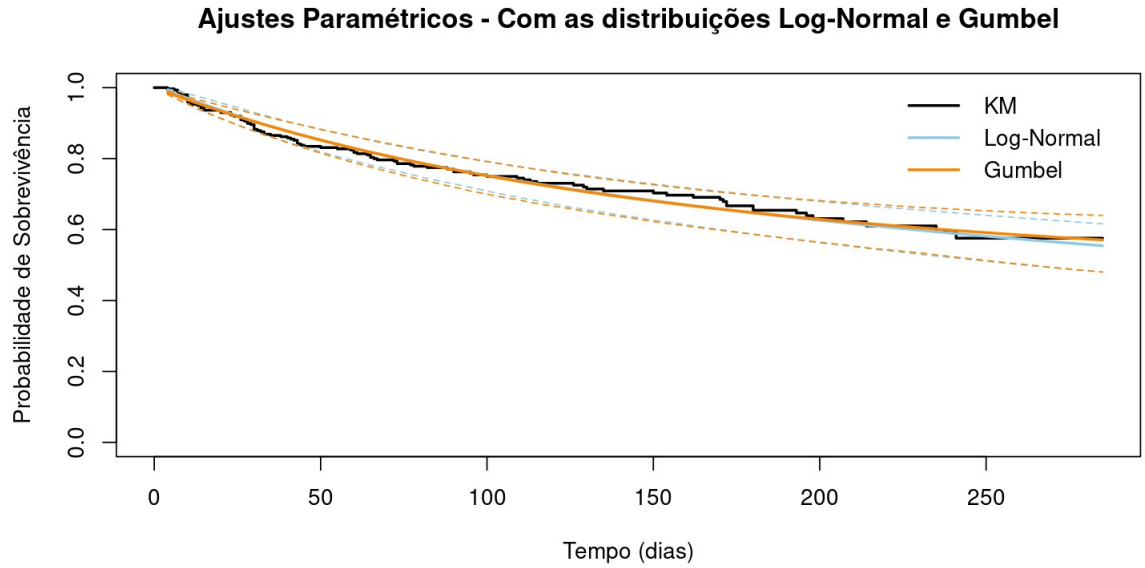


Tabela 8: Critério de Informação de Akaike (AIC) para diferentes distribuições ajustadas

Modelo	**df**	**AIC**
Exponencial	1	1347.082
Weibull	2	1344.876
Log-Normal	2	1336.546
Gumbel	2	1341.495

Com base na Tabela 8 de critérios de informação (AIC), observou-se que a distribuição Lognormal apresentou o menor AIC, confirmando o melhor ajuste, em consonância com a análise gráfica. Contudo, considerando a proximidade entre os valores de AIC das

distribuições Lognormal e Gumbel, Nesse caso o Teste de Razão de verossimilhança(LTR) para comparação entre os modelos não é útil pois a distribuição gumbel não faz parte da mesma família de distribuição da Log-normal.

Durante a etapa de ajuste dos modelos de mistura paramétricos, ocorreram dificuldades na implementação via pacote flexsurvmix do R, que não reconheceu as variáveis essenciais à análise de sobrevivência (tempo dias e Evento de morte). Diante disso, o procedimento foi migrado para o Google Colab, utilizando as bibliotecas NumPy, Pandas e SciPy. Apesar de o modelo Log-Normal ter apresentado o melhor ajuste para o modelo de componente único ($k = 1$), essa abordagem não é suficiente no caso de KM, pois quando temos observações que não experimentaram o evento de interesse, nesse cenário a curva de KM se estabilizará em uma proporção diferente de zero. Logo o modelo de cura AFT podem levar em conta efeitos de curto e longo prazo e ser calculados separadamente.

Essa característica constitui uma evidência visual e estatística de heterogeneidade na amostra, indicando que o uso de um modelo com $k = 1$ pode mascarar a existência de subgrupos com taxas de risco claramente distintas. Para lidar com essa heterogeneidade, torna-se necessária a utilização de um modelo de mistura de distribuição AFT, o qual permite representar a população como uma soma ponderada de k distribuições, cada uma correspondendo a um subgrupo latente com comportamento próprio de risco.

13.9 Modelos de Mistura AFT com Diferentes Distribuições

Com o objetivo de capturar a heterogeneidade existente na amostra de pacientes com insuficiência cardíaca, foram ajustados modelos de mistura Acelerados de Tempo de Falha (AFT) considerando Weibull, Gumbel e Lognormal, todos com o número de componentes variando entre $K = 2$ e $K = 3$. Em todos os ajustes, utilizou-se a variável *Fração de Ejeção* (FE) como principal covariável explicativa, por ser um importante marcador prognóstico de risco e para observar se ela ajuda a indicar a subgrupos naturais que podem estar associados aos grupos latentes.

13.9.1 Modelo de Mistura Weibull AFT

O modelo de mistura Weibull com $K = 2$ componentes apresentou log-verossimilhança de -649.93, com critérios de informação $AIC = 1313.86$ e $BIC = 1339.77$, enquanto o modelo com $K = 3$ componentes resultou em $AIC = 1300.28$ e $BIC = 1340.99$. Assim, de acordo com o critério de Bayes (BIC), o modelo com $K = 2$ foi selecionado por apresentar melhor equilíbrio entre ajuste e parcimônia.

A distribuição dos pacientes entre os dois componentes indicou que o cluster 0 (n=279) representa a maioria dos pacientes, enquanto o cluster 1 (n=20) corresponde a um subgrupo menor. A média da fração de ejeção foi de aproximadamente $37,6\%$ no cluster 0 e $44,1\%$ no cluster 1. A interpretação clínica sugere que o componente menor reúne pacientes com função cardíaca relativamente preservada (fração de ejeção mais alta), enquanto o grupo maior concentra indivíduos com disfunção ventricular mais pronunciada, refletindo maior gravidade da insuficiência cardíaca.

13.9.2 Modelo de Mistura Gumbel AFT

O modelo de mistura Gumbel com $K = 2$ apresentou log-verossimilhança de -521.34, sendo selecionado também pelo menor valor de $AIC = 1056,68$ $BIC = 1082,59$. A divisão amostral resultou em 271 pacientes no cluster 0 e 28 no cluster 1. Observou-se que o cluster 1 apresentou *Fração de Ejeção* média inferior ($34,2\%$) em comparação ao cluster 0 ($38,5\%$), além de maior média de creatinina sérica ($2,24$ contra $1,30$) e maior prevalência de anemia. Esse padrão reforça que o componente 1 agrupa indivíduos em condição clínica mais crítica, caracterizada por pior função renal e função cardíaca mais comprometida, com isso o modelo gumbel apresentou o melhor desempenho dos modelos em todos os critérios de desempenho global.

13.9.3 Modelo de Mistura Lognormal AFT

Entre todas as distribuições analisadas, o modelo de mistura Lognormal com $K = 2$ apresentou o segundo melhor desempenho global, com log-verossimilhança de -640.95, $AIC = 1295.90$ e $BIC = 1321.80$, superando os modelos Weibull em ambos

os critérios de informação. Esse resultado indica que a distribuição lognormal oferece o segundo melhor ajuste aos tempos de sobrevida observados, preservando a interpretação paramétrica da taxa de falha.

A análise da composição dos clusters mostrou que o cluster 0 (n=237) apresenta pacientes com fração de ejeção média de $37,8\%$, idade média de $58,9$ anos e creatinina sérica média de $1,24$, enquanto o cluster 1 (n=62) reúne pacientes com fração de ejeção ligeiramente maior ($39,1\%$), idade média superior ($68,2$ anos) e níveis mais elevados de creatinina sérica ($1,97$). Essa estrutura sugere a presença de dois subgrupos distintos: um composto majoritariamente por pacientes mais jovens com perfil hemodinâmico estável, e outro por indivíduos mais idosos e com comprometimento renal relevante, mesmo apresentando fração de ejeção apenas discretamente superior.

13.9.4 Síntese Comparativa

A Tabela 9 resume os resultados comparativos dos modelos ajustados. Considerando os valores de log-verossimilhança, AIC e BIC, o modelo Gumbel AFT com $K = 2$ componentes foi considerado o mais parcimonioso e adequadamente ajustado aos dados, sendo, portanto, selecionado como modelo final para as análises subsequentes de sobrevivência.

Tabela 9: Comparação dos modelos de mistura AFT ajustados

Modelo	**K**	**LogLik**	**AIC**	**BIC**
Weibull AFT	2	-649.93	1313.87	1339.77
Gumbel AFT	2	-521.34	1056.68	1082.59
Lognormal AFT	2	-640.95	1295.90	1321.81

14 Análise dos gráficos

14.1 Análise Visual da Separação dos Clusters (Responsabilidade)

Uma etapa crucial na seleção do modelo de mistura aft é avaliar a certeza com que o modelo atribui cada paciente a um cluster. Isso é feito analisando a Responsabilidade(probabilidade de pertencimento) ao Componente 0, onde valores próximos de 1 ou 0 indicam alta certeza, e valores intermediários (0,2-0,8) indicam incerteza (separação "nebulosa"ou *fuzzy*), pacientes perto de 0 são os ditos "curados"(Bola azul) e os perto de 1 são os suscetiveis ou seja pacientes com alto risco(Bola laranja)(KUO; PENG, 1995).

14.2 Modelo de Mistura Gumbel AFT (K=2)

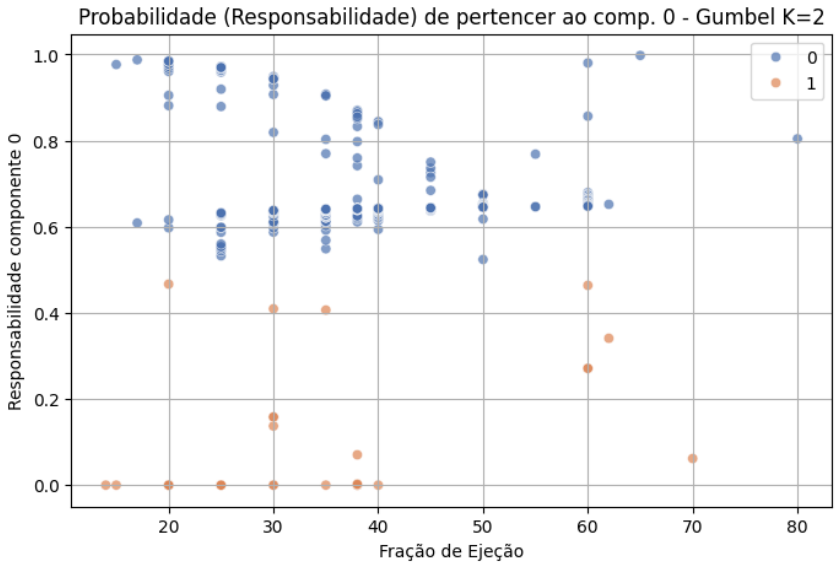


Figura 1: Responsabilidade vs. FE (Gumbel AFT K=2)

O modelo Gumbel (Figura 1) demonstrou a separação mais fraca e nebulosa dos clusters. Observa-se uma incerteza considerável, com muitos pacientes apresentando probabilidades de responsabilidade na faixa intermediária (entre 0,2 e 0,8). Isso sugere que a distribuição Gumbel não foi capaz de encontrar uma fronteira estatística clara para definir os dois subgrupos de risco.

14.3 Modelo de Mistura Weibull AFT (K=2)

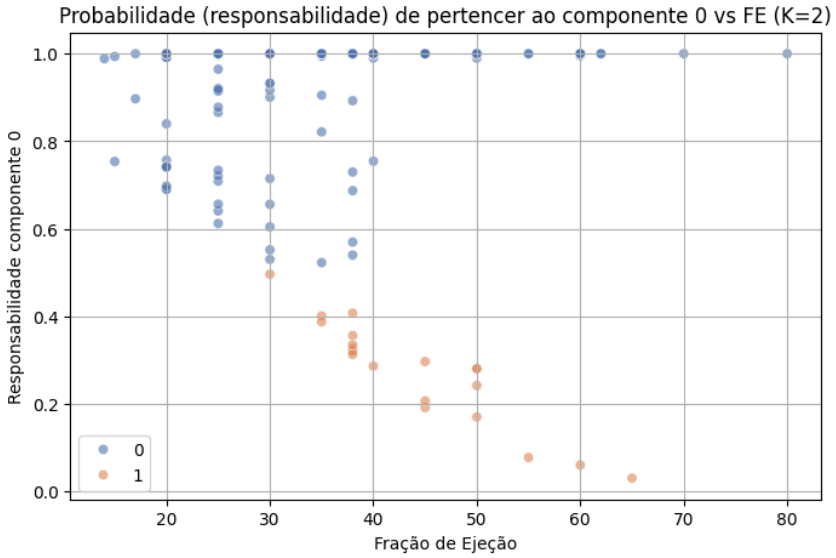


Figura 2: Responsabilidade vs. FE (Weibull AFT K=2)

O modelo Weibull (Figura 2) apresentou uma separação muito superior à do Gumbel, com um agrupamento mais "duro". A maioria dos pontos está polarizada nos extremos (próximo de 1,0 para o Cluster 0 e próximo de 0,0 para o Cluster 1). No entanto, ainda existe uma zona de transição, principalmente na faixa de FE entre 35-50, onde alguns pacientes demonstram incerteza.

14.4 Modelo de Mistura Lognormal AFT (K=2)

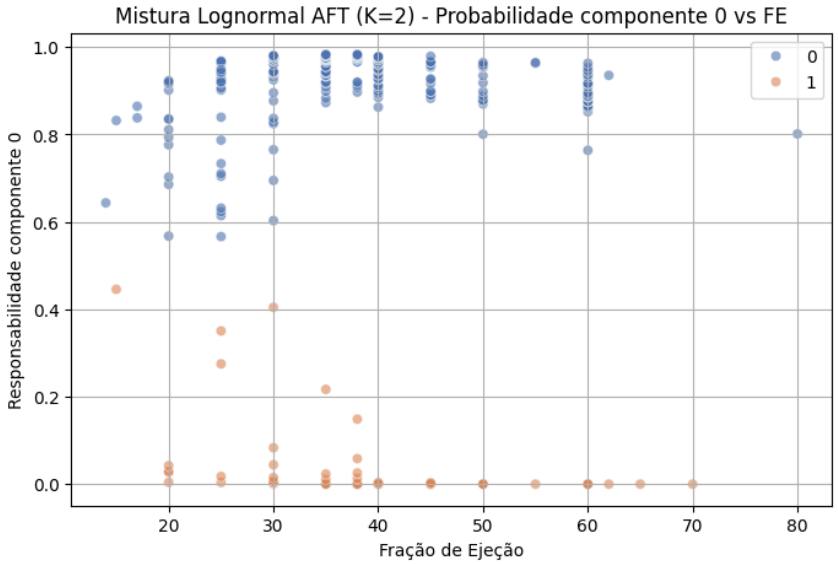


Figura 3: Responsabilidade vs. FE (Lognormal AFT K=2)

O modelo Lognormal (Figura 3) demonstrou a separação dos clusters mais limpa e distinta entre os três modelos. A separação é quase perfeitamente polarizada: quase todos os pacientes do Cluster 0 (azul) têm responsabilidade $\approx 1,0$, e quase todos do Cluster 1 (laranja) têm responsabilidade ≈ 0.

14.5 Conclusão Comparativa Visual

A comparação visual das Figuras 1, 2 e 3 O fato da distribuição gumbel ter tido os melhores critérios de sempenho a distribuição de Mistura Lognormal AFT (K=2) teve o melhor ajuste de clusters, classificando os pacientes com menor ambiguidade.

15 Conclusão

A etapa final da aplicação consistiu em validar a capacidade do Modelo de Mistura AFT Gumbel ($K = 2$) de estratificar o experimento em subgrupos de risco clinicamente significativos. A validação foi realizada através da análise visual de Kaplan-Meier e da significância estatística do Teste Log-Rank.

15.1 Validação Visual por Kaplan-Meier

A Figura 4 apresenta as curvas de sobrevivência de Kaplan-Meier para os dois clusters identificados pelo modelo AFT Gumbel.

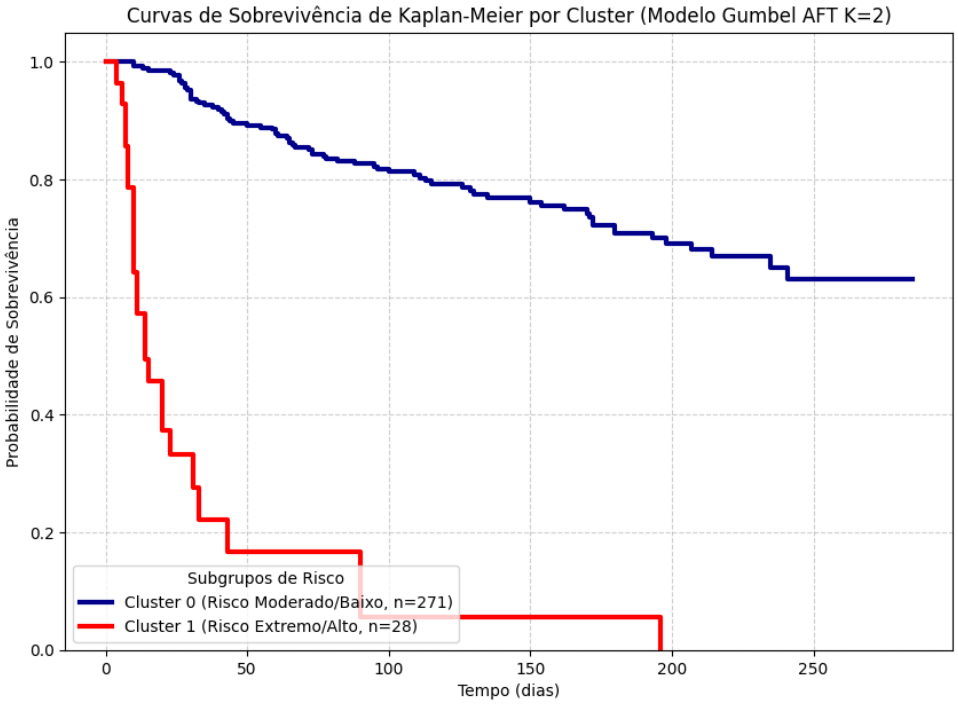


Figura 4: Curvas de Sobrevivência de Kaplan-Meier por Cluster (Modelo Gumbel AFT K=2)

- **Cluster 0 (Risco Moderado/Baixo, $n = 271$):** Este subgrupo (Curva Azul)já possui um prognóstico melhor, a curva de sobrevivência é relativamente estável, mantendo acima de 65% memos após 250 dias de acompanhamento e isso significa que esses pacientes possuem uma condição menos grave.

- **Cluster 1 (Risco Extremo/Alto, $n = 28$):** A curva Vermelha demonstra que esse subgrupo apresenta uma taxa de falha extremamente alta e precoce, a probabilidade de sobrevivência cai abaixo de 50% em menos de 10 dias e a maioria dos eventos de falha ocorre antes dos 100 dias.

15.2 Validação Estatística por Teste Log-Rank

Para quantificar a diferença observada na Figura 4, foi aplicado o Teste Log-Rank, cujos resultados são apresentados na Tabela 10.

Tabela 10: Resultado do Teste Log-Rank para os Clusters do Modelo gumbel K=2

Métrica	Valor
Estatística do Teste (χ^2)	$194,07$
Graus de Liberdade (gl)	1
Valor-p	$< 0,005$

O teste resultou em um valor de $\chi^2 = 194,07$ com um valor-p altamente significativo ($p < 0,005$). Este resultado rejeita categoricamente a hipótese nula de que as curvas de sobrevivência dos dois clusters são idênticas.

15.3 Conclusão da Aplicação

A combinação da separação visual extrema (Figura 4) e a validação estatística robusta (Tabela 10) confirma que o Modelo de Mistura AFT Gumbel ($K = 2$) foi altamente bem-sucedido em identificar a heterogeneidade do experimento. O modelo estratificou a população em dois subgrupos com prognósticos de mortalidade fundamentalmente distintos, validando a metodologia como uma ferramenta superior para a descoberta de perfis de risco latentes.

16 Apêndice

Aplicação da Desigualdade de Jensen

A variação da log-verossimilhança é dada por:

$$l(\Psi^{(s)}|\mathcal{X}) - l(\Psi^{(s-1)}|\mathcal{X}) = [Q(\Psi^{(s)};\Psi^{(s-1)}) - Q(\Psi^{(s-1)};\Psi^{(s-1)})] + [H(\Psi^{(s)};\Psi^{(s-1)}) - H(\Psi^{(s-1)};\Psi^{(s-1)})]$$

é definido a diferença entre os termos H como:

$$\Delta H = \int p(\mathcal{Z}|\mathcal{X},\Psi^{(s-1)}) \ln p(\mathcal{Z}|\mathcal{X},\Psi^{(s-1)})\, d\mathcal{Z} - \int p(\mathcal{Z}|\mathcal{X},\Psi^{(s-1)}) \ln p(\mathcal{Z}|\mathcal{X},\Psi^{(s)})\, d\mathcal{Z}$$

Juntando os termos:

$$\Delta H = \int p(\mathcal{Z}|\mathcal{X}, \Psi^{(s-1)}) \ln \left(\frac{p(\mathcal{Z}|\mathcal{X}, \Psi^{(s-1)})}{p(\mathcal{Z}|\mathcal{X}, \Psi^{(s)})} \right) d\mathcal{Z}$$

Aplicando a Desigualdade de Jensen para a função convexa $-\ln(x)$:

$$-\ln \left(\int p_{s-1} \frac{p_s}{p_{s-1}} \, d\mathcal{Z} \right) \leq \int p_{s-1} \left(-\ln \frac{p_s}{p_{s-1}} \right) d\mathcal{Z}$$

Com a integral do lado esquerdo é 0

$$\int p_{s-1} \frac{p_s}{p_{s-1}} \, d\mathcal{Z} = \int p_s \, d\mathcal{Z} = 1$$

temos a prova da não-negatividade da Divergência KL:

$$0 \leq \int p_{s-1} \ln \left(\frac{p_{s-1}}{p_s} \right) d\mathcal{Z}$$

Referências

AITKIN, M.; CLAYTON, D. The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1980.

ANNUAL Review of Statistics and Its Application. Finite Mixture Models. 2019.

APLICAÇÕES DE CÓPULAS EM MODELOS DE RISCOS MÚLTIPLOS DEPENDENTES E EM MODELOS DE MISTURAS DE DISTRIBUIÇÕES. 2012.

AZEVEDO, C. Modelos mistos (análise de medidas repetidas).

CHICCO, D.; JURMAN, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, London, 2020.

COLOSIMO, E. A. Análise de Sobrevivência Aplicada. 2024.

COLOSIMO, E. A. Métodos Estatísticos Avançados em Epidemiologia: Análise de Sobrevivência – Modelo de Cox. Departamento de Estatística, Universidade Federal de Minas Gerais.

COUTO, E. V. de. Modelo de regressão log-gama generalizado exponenciado com dados censurados. 2010.

COX, D. R. Partial likelihood. Biometrika, 1975.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm.Journal of the Royal Statistical Society. Series B (Methodological), 1977.

Ding, A. A., & Nan, B. A semiparametric model for clustered survival data with informative cluster sizes. Biostatistics, 2015.

Gaussian Processes for Machine Learning Autores: Carl Edward Rasmussen & Christopher K. I. Williams Editora: MIT Press, 2006.

Kalbfleisch, J. D., & Prentice, R. L. (1980). The Statistical Analysis of Failure Time Data. Wiley.

KRACHEY, E. C. Variations on the accelerated failure time model: mixture distributions, cure rates, and different censoring scenarios, 2009.

KUO, L.; PENG, F. A Mixture-Model Approach to the Analysis of Survival Data. In: SALAHUDDIN, S. (ed.). Bayesian Statistics and Its Applications. New York: Springer, 1995.

MASSUIA, M. B. Modelos para dados censurados sob a classe de distribuições misturas de escala skew-normal. 2015.

MCLACHLAN, G.; PEEL, D. Finite Mixture Models. New York: Wiley, 2000.

MELO, V. S. Modelos Envolvendo Duas Distribuições Weibull Inversa e Apresentação da Weibull Inversa Generalizada. 2014.

MOURA, I. C. G. Diagnóstico de influência em modelos de regressão para dados censurados utilizando distribuições de caudas pesadas. 2016.

OLIVEIRA, Márcio Vinícius de. Análise de desigualdade educacional: comparação entre uma divergência Kullback-Leibler contínua e uma degenerada discreta. 2017.

PAULA, G. A. Modelos de regressão com apoio computacional. São Paulo: Instituto de Matemática e Estatística da Universidade de São Paulo, [s.d.]. Disponível em: e-mail: giapaula@ime.usp.br.

PESCIM, R. R. Análise de Sobrevivência. [Playlist]. YouTube. Londrina, 2024. Disponível em: [https://youtu.be/FX05fvZlx5I?si=7oCdzDUayCGxhywA].

PHILOSOPHICAL Transactions of the Royal Society of London. A. Vol. 185. 1894.

ROHDE, L. E. P. et al. Diretriz Brasileira de Insuficiência Cardíaca Crônica e Aguda. Arquivos Brasileiros de Cardiologia, Rio de Janeiro, 2018.

ROSA, M. V. F. Modelos de Misturas Gaussianas.Florianópolis, 2022.

SANTOS, I. de S.; BITTENCOURT, M. S. Insuficiência cardíaca. Revista de Medicina (São Paulo), São Paulo,2008.

SILVA, Y. C. Sobre Identificabilidade para Misturas Finitas. Brasília, 2013.

SOUZA, F. H. R.; PINTO, E. J. A. Tratamento de dados censurados de nitrogênio

amoniacal total pelos métodos de máxima verossimilhança, robustos e Kaplan-Meier. In: SIMPÓSIO DE RECURSOS HÍDRICOS DO NORDESTE.

SOUZA, F. L. de. Mistura de Distribuições Extremais. Maio de 2010.

SOUZA, F. L. de. LARA DE SOUZA, F. Tratamento de dados censurados em estudos ambientais. Janeiro de 2014.

SOUZA SANTOS, D. de. Comparações múltiplas para dados censurados. 2013. Dissertação (Mestrado) – Universidade de São Paulo, São Carlos.

Susarla, V., & Van Ryzin, J. Nonparametric Bayesian estimation of survival curves from incomplete observations. Journal of the American Statistical Association, 1976.

TACONELI, J. P. Modelo de mistura paramétrico com fragilidade na presença de covariáveis. Junho de 2013.

TITTERINGTON, D. M.; SMITH, A. F. M.; MAKOV, U. E. Statistical Analysis of Finite Mixture Distributions. New York: Wiley, 1985.

WALD, A. Note on the consistency of the maximum likelihood estimate. Annals of Mathematical Statistics, 1949.

Wei, G. C. G., & Tanner, M. A. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association, 1990.

ZÖRNIG, P. Probability Theory and Statistics with Real World Applications: Univariate and Multivariate Applications. Berlin; Boston: De Gruyter, 2024.