Universidade Federal De Juiz de Fora
Instituto de Ciências Exatas/Faculdade de Engenharia
Programa de Pós-Graduação em Modelagem Computacional

Carlos Alberto Huaira Contreras

Um modelo adaptativo de sobrevivência baseado em Buckley-James e Comitê L2 Boosting de Máquinas de Aprendizado Extremo, com abordagem automática para determinação de não linearidade e robustez a observações extremas

Juiz de Fora
2025

Carlos Alberto Huaira Contreras

Um modelo adaptativo de sobrevivência baseado em Buckley-James e Comitê L2 Boosting de Máquinas de Aprendizado Extremo, com abordagem automática para determinação de não linearidade e robustez a observações extremas

Tese apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Doutor em Modelagem Computacional.

Orientador: Prof. D.Sc Carlos Cristiano Hasenclever Borges

Juiz de Fora

2025

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Huaira Contreras, Carlos Alberto.
Um modelo adaptativo de sobrevivência baseado em Buckley-James e Comitê L2 Boosting de Máquinas de Aprendizado Extremo, com abordagem automática para determinação de não linearidade e robustez a observações extremas / Carlos Alberto Huaira Contreras. -- 2025.
104 f. : il.

Orientador: Carlos Cristiano Hasenclever Borges
Tese (doutorado) - Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2025.

1. Análise de Sobrevivência. 2. Modelo de Buckley-James. 3. Máquina de Aprendizado Extremo. 4. Boosting L2. 5. Distribuição t-Student. I. Hasenclever Borges, Carlos Cristiano, orient. II. Título.

Carlos Alberto Huaira Contreras

Um Modelo Adaptativo de Sobrevivência baseado em Buckley-James e Comitê L2 Boost de Máquinas de Aprendizado Extremo, com Abordagem Automática para Determinação de Não Linearidade e Robustez a Observações Extremas

Tese apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutor em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Aprovada em 17 de dezembro de 2025.

BANCA EXAMINADORA

Prof. Dr. Carlos Cristiano Hasenclever Borges - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Leonardo Goliatt da Fonseca
Universidade Federal de Juiz de Fora

Prof.ª Dr.ª. Camila Borelli Zeller
Universidade Federal de Juiz de Fora

Prof. Dr. Fabrízzio Condé de Oliveira
Universidade Federal Fluminense

Dr.ª Rossana Verónica Mendoza López
Universidade de São Paulo

Juiz de Fora, 09/12/2025.

Documento assinado eletronicamente por **Carlos Cristiano Hasenclever Borges**, **Professor(a)**, em 17/12/2025, às 14:54, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Camila Borelli Zeller**, **Professor(a)**, em 18/12/2025, às 09:31, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rossana Veronica Mendoza Lopez**, **Usuário Externo**, em 19/12/2025, às 09:54, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fabrízzio Condé de Oliveira**, **Usuário Externo**, em 19/12/2025, às 12:56, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca**, **Professor(a)**, em 09/02/2026, às 19:46, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

AGRADECIMENTOS

Ao Programa de Pós-Graduação em Modelagem Computacional (PPGMC) da Universidade Federal de Juiz de Fora (UFJF), pela oportunidade concedida. Aos amigos, professores e funcionários, pelo acompanhamento ao longo desta trajetória. À UFJF e à CAPES, pelo suporte financeiro indispensável ao desenvolvimento deste trabalho. Um agradecimento especial ao meu orientador, Prof. Dr. Carlos Cristiano Hasenclever Borges, por sua dedicação, colaboração, ensinamentos e pela confiança depositada.

Aos membros da banca examinadora, pelas valiosas contribuições oferecidas a este trabalho. À Professora Dra. Natália Maria da Silva Fernandes, pela disponibilização do conjunto de dados e pela pronta disposição em esclarecer e sanar eventuais dúvidas. Ao João Víctor Lauro e à Rosália Maria Nunes Henriques Huaira, cuja assistência no manejo dos dados e apoio na programação foram determinantes para a realização desta tese de doutorado.

Finalmente, agradeço à minha família, tanto brasileira quanto peruana, em especial à minha esposa Rosália e ao meu filho Gael, por me acompanharem e apoiarem em todos os momentos. À memória de Ana Maria e Albertina, cuja presença espiritual e inspiração me sustentaram ao longo desta jornada.

RESUMO

Os modelos de sobrevivência são amplamente utilizados em estudos que analisam o tempo até a ocorrência de um evento de interesse. Esses modelos lidam com conjuntos de dados onde, para algumas observações, o evento pode não ocorrer durante o período de acompanhamento, resultando em dados censurados, nos quais o tempo até o evento não é completamente conhecido. Tradicionalmente, tais modelos têm sido aplicados na área médica para avaliação de tempos de vida, de recidiva de doenças e na modelagem de problemas de confiabilidade como, por exemplo, para determinação de tempos de falhas em componentes mecânicos ou eletrônicos. Mais recentemente, também têm encontrado aplicações em ciências sociais na modelagem de problemas relacionados à internet, como a determinação do tempo até o abandono de um emprego, a avaliação do período para saída de plataformas digitais e, até mesmo, na determinação temporal para conclusão de sessões de navegação. Neste contexto, o modelo semiparamétrico de Buckley-James (BJ) surge como alternativa. Este pode ser visto como uma extensão da regressão linear para situações com censura permitindo, assim, o uso de mínimos quadrados ordinários nos processos de estimação. Diferentemente do modelo de Cox, o BJ não exige a suposição de proporcionalidade dos riscos. No entanto, a eficiência do modelo BJ é comprometida quando a relação entre covariáveis e resposta não é linear e/ou quando há presença de observações extremas (*outliers*). Este trabalho propõe uma abordagem computacional adaptativa para modelos de sobrevivência com censura à direita, construída no âmbito do modelo BJ. A proposta integra em um processo de aprendizagem um comitê de regressores baseado nas Máquinas de Aprendizado Extremo, que são redes neurais conhecidas pela sua eficiência computacional em um esquema L2 Boosting com ponderação de casos baseada na distribuição *t-Student*. Essa formulação substitui a combinação linear de covariáveis do modelo BJ pela função de saída do comitê, permitindo que o modelo desenvolvido selecione automaticamente entre estruturas lineares e não lineares e, adicionalmente, incorpore robustez na presença de observações extremas. A escolha da função de ativação nos ELM possibilita capturar diferentes padrões de relação entre covariáveis e resposta, enquanto os graus de liberdade da distribuição *t-Student* controlam a sensibilidade do modelo a valores extremos. Os resultados obtidos a partir de experimentos com dados simulados e com conjuntos de dados referenciados na literatura, avaliados por métricas como C-Index e IBS, evidenciam ganhos significativos de flexibilidade e desempenho, constituindo a base para o desenvolvimento de um método geral e unificado. Adicionalmente, a proposta é aplicada a um conjunto de dados de pacientes brasileiros com doença renal crônica, demonstrando sua relevância prática e o potencial para análises em cenários clínicos reais.

Palavras-chave: Análise de Sobrevivência. Modelo de Buckley-James. Máquina de Aprendizado Extremo. Boosting L2. Distribuição t-Student.

ABSTRACT

Survival models are widely used in studies that analyze the time until the occurrence of an event of interest. These models deal with datasets in which, for some observations, the event may not occur during the follow-up period, resulting in censored data where the time to the event is not fully known. Traditionally, such models have been applied in the medical field for the evaluation of lifetimes, disease recurrence, and in the modeling of reliability problems, such as determining failure times in mechanical or electronic components. More recently, they have also found applications in the social sciences for modeling problems related to the internet, such as determining the time until job abandonment, evaluating the period until leaving digital platforms, and even determining the time to conclude browsing sessions.

In this context, the semiparametric Buckley-James (BJ) model emerges as an alternative. It can be seen as an extension of linear regression to censored situations, thus allowing the use of ordinary least squares in estimation processes. Unlike the Cox model, BJ does not require the assumption of proportional hazards. However, the efficiency of the BJ model is compromised when the relationship between covariates and response is not linear and/or when extreme observations (outliers) are present.

This work proposes an adaptive computational approach for right-censored survival models, built within the BJ framework. The proposal integrates into a learning process a committee of regressors based on Extreme Learning Machines, which are neural networks known for their computational efficiency, in an L2 Boosting scheme with case weighting based on the Student's t-distribution. This formulation replaces the linear combination of covariates in the BJ model with the output function of the committee, allowing the developed model to automatically select between linear and nonlinear structures and, additionally, incorporate robustness in the presence of extreme observations. The choice of activation function in the ELM enables capturing different patterns of relationships between covariates and response, while the degrees of freedom of the Student's t-distribution control the model's sensitivity to extreme values.

The results obtained from experiments with simulated data and datasets referenced in the literature, evaluated by metrics such as C-Index and IBS, demonstrate significant gains in flexibility and performance, forming the basis for the development of a general and unified method. Additionally, the proposal is applied to a dataset of Brazilian patients with chronic kidney disease, demonstrating its practical relevance and potential for analyses in real clinical scenarios.

Keywords: Survival Analysis. Buckley-James model. Extreme Learning Machine. boosting L2. t-Student distribution.

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO DO PROBLEMA

Desde 1662, ano em que John Graunt apresentou alguns resultados descritivos sobre números de nascimentos e mortes e a relação destes com certas doenças na Inglaterra, utilizando um formato rudimentar de tabelas de vida [1], o interesse por estudar as populações para avaliar a natalidade, mortalidade e os padrões de fecundidade, associados a alguns atributos e eventos tornou-se importante.

Graunt realizou um trabalho puramente empírico, avaliando dados reais registrados em paróquias e obtidos por buscadoras. A partir desses dados, construiu as tabelas de vida considerando alguns intervalos de idade e calculando taxas de sobrevivência que foram associadas às causas de morte registradas. Considera-se que este trabalho mostrou a importância da demografia e da ciência atuarial, além de ter estabelecido as bases para o que conhecemos atualmente como epidemiologia.

Durante os três séculos seguintes, a pouca disponibilidade de dados, a falta de precisão e sistematização na coleta destes, e a ausência de dados obtidos por procedimentos experimentais e planejados, levaram à obtenção de conclusões empíricas e descritivas sobre o comportamento das populações. Durante este período, propuseram-se funções teóricas explícitas para descrever a mortalidade populacional [2], sendo duas das mais conhecidas a aproximação linear de De Moivre (1725) e a aproximação de Gompertz (1825).

Após o fim da segunda guerra mundial, o grande interesse por duas áreas específicas tornou-se fundamental para o desenvolvimento e formalização do que é conhecido como Análise de Sobrevivência, em inglês *Survival Analysis*. Por um lado, havia o interesse no melhoramento da saúde e da medicina preventiva, e, por outro, o interesse na melhoria da qualidade e confiabilidade de produtos e serviços. Para esse fim, foram introduzidas diversas técnicas estatísticas de análise de dados para esses tipos de situações.

Na área da saúde, a busca por respostas a muitos problemas epidemiológicos levou à realização de diversas pesquisas. Um interesse especial foi dado a estudos sobre câncer. A identificação de fatores de risco (comportamentais, hábitos e ambientais, entre outros) e a avaliação do tempo de sobrevida de indivíduos que possuem caraterísticas formadas por uma combinação dos fatores avaliados, assim como a comparação dos tempos de sobrevida entre grupos submetidos a determinado tratamento e um grupo controle, tornaram-se frequentes. Além disso, ensaios e testes para diversas vacinas desenvolvidas foram frequentes neste período. Como exemplo, podemos citar o estudo que gerou a publicação de um relatório emblemático que associava o tabagismo ao câncer de pulmão em 1964 [3].

Por outro lado, a busca contínua pela qualidade levou ao surgimento da denominada engenharia de confiabilidade, a qual faz uso de diversos métodos estatísticos para

alcançar seus objetivos. Ao definir a confiabilidade como a qualidade de um produto ao longo do tempo, todos os envolvidos nestes processos consideram que essa caraterística é indispensável para competir nos mercados atuais. Considerando uma definição mais formal da confiabilidade como sendo a probabilidade de uma unidade desempenhar sua função até o tempo especificado, sob as condições de uso encontradas [4], foram propostos modelos estatísticos para estimar os tempos de desempenho, inicialmente usando a distribuição exponencial e posteriormente distribuições Weibull e Lognormal [5].

Em anos recentes, a Análise de Sobrevivência foi aplicada em outras áreas de pesquisa. Dirick (2017) [6] apresenta uma visão geral do uso dessas análises na modelagem de risco de crédito, uma área com grande interesse e potencial de pesquisa nos dias atuais. Uma revisão que descreve diversos trabalhos com aplicações da Análise de Sobrevivência em diversas áreas é apresentado por Wang (2019) [7]. Neste trabalho são mencionadas aplicações na área social, onde se avaliam o tempo para reinserção no mercado de trabalho, explicado por caraterísticas sociodemográficas do indivíduo e indicadores econômicos, e o tempo de desistência escolar, explicado por caraterísticas demográficas, financeiras e acadêmicas do aluno. O trabalho também apresenta aplicações associadas ao uso de internet que em anos recentes tomou grande relevância. Uma primeira aplicação avalia o tempo para a compra de um determinado serviço, explicado por dados sociodemográficos e caraterísticas do produto e da loja virtual; uma segunda aplicação avalia o tempo para o sucesso de um projeto em financiamentos colaborativos (em inglês, *Crowdfunding*) explicado por caraterísticas do projeto, dos criadores e de caraterísticas comunicacionais via Twiter; e, numa terceira aplicação, avalia-se o tempo para visualização de propaganda explicado por caraterísticas demográficas e interesses do usuário, bem como caraterísticas da propaganda na página eletrônica. Considera-se que estudos desse tipo têm grande potencial futuro, devido aos constantes e rápidos avanços no uso da internet e à enorme quantidade de informação que é produzida e possível de ser analisada. Finalmente, são descritas aplicações recentes e cada vez mais frequentes em ciências da saúde e bioinformática, que incluem informações de expressão genética, levando a análises com uma grande quantidade de atributos (em inglês, *high dimensional data*).

Embora as técnicas estatísticas utilizadas na Análise de Sobrevivência sejam muito conhecidas e estabelecidas, elas são o ponto de partida para o desenvolvimento de modelos que tentam lidar de forma mais eficaz com as novas situações trazidas pela disponibilidade de uma grande quantidade de informações e pelos avanços computacionais recentes. Por um lado, a disponibilidade de grande quantidade de informação conduz a três questões que devem ser consideradas: o modelo deve ser capaz de lidar com uma grande quantidade de atributos disponíveis para a explicação de um fenômeno (*high dimensional*), as relações não lineares presentes na estrutura do modelo e ser robusto em relação à presença de dados extremos (*outliers*) dentro do conjunto de informações disponível. Os avanços na área computacional, por sua vez, permitem a criação de modelos estatísticos estruturalmente

mais complexos e com novas abordagens probabilísticas, descrevendo com maior fidelidade os fenômenos investigados.

Nesta linha de pesquisa, algoritmos de Aprendizado de Máquina são propostos para a solução de problemas em Análise de Sobrevivência, demostrando grande desempenho quando avaliados por seus resultados preditivos. Diversos estudos que combinam Análise de Sobrevivência com algoritmos de Aprendizado de Máquina para obter melhores resultados na predição do tempo até a ocorrência do evento vem sendo desenvolvidos recentemente (Wang, 2019 [7]).

1.2 MOTIVAÇÃO

O trabalho é motivado por um problema real relacionado a dados médicos associados ao tratamento de Doença Renal Crônica (DRC). A questão principal é propor um algoritmo computacional para a obtenção de previsões confiáveis do tempo de sobrevida (em meses) de pacientes com DRC levando em consideração um conjunto de características dos pacientes.

Desta forma, além do registro do tempo de sobrevida do paciente com DRC, o banco de dados contém variáveis demográficas, médicas, laboratoriais, de avaliação clínica e de qualidade de vida. A ocorrência de heterogeneidade das informações provenientes de diversas fontes e as relações entre as variáveis medidas indicam que o modelo proposto deve lidar com estruturas mais complexas para descrever o fenômeno estudado. Isso inclui um alto número de atributos, relações não lineares entre os tempos de sobrevida e os atributos considerados, e a possibilidade de presença de dados extremos. Neste contexto, a utilização de um algoritmo de Aprendizado de Máquinas torna-se uma boa alternativa para lidar com este conjunto de situações.

A busca por algoritmos de Aprendizado de Máquina para a solução de problemas de dados médicos, como o descrito acima, conduziu à constatação de duas situações recorrentes na atualidade. Primeiro, a crescente utilização da Análise de Sobrevivência em diversos contextos (veja por exemplo, Dirick (2017) [6] e Wang (2019) [7]). Pode-se encontrar aplicações que ilustram a versatilidade da Análise de Sobrevivência ao tratar de tempos até eventos em áreas tão diversas quanto saúde, economia e comportamento social. Segundo, diversos algoritmos de Aprendizado de Máquina têm sido propostos para lidar com as características complexas dos dados, como alta dimensionalidade, não linearidade e a presença de *outliers*. Por exemplo, técnicas como Redes Neurais e Florestas Randômicas têm mostrado melhor desempenho em contextos de alta dimensionalidade e não linearidade (Ishiwaran, 2007 [8] e Zhao et al., 2019 [9]), enquanto métodos baseados em Boosting e Máquinas de Vetores Suporte são frequentemente utilizados para garantir robustez na presença de dados extremos (Chen et al., 2016 [10] e Zhu et al. [11]).

Observa-se que todas as propostas e avanços apresentados refletem a crescente importância de adaptar os algoritmos às especificidades dos dados, garantindo maior

precisão nas previsões. Assim, antes da resolução do problema que motiva o trabalho decidiu-se que em primeiro lugar deveria se propor um algoritmo de Aprendizado de Máquina que possa ser utilizado em diversos problemas de Análise de Sobrevivência e que sejam flexíveis para incorporar (considerar ou incluir) na análise as especificidades do problema que está sendo estudado.

Nos últimos anos, a Máquina de Aprendizado Extremo (ELM) (Huang et al., 2004 ; 2006 [12] [13]) tem sido utilizada para diferentes tarefas de aprendizado. Wang et al. [14] apresentam uma revisão sobre o ELM indicando seu uso em tarefas de aprendizado para classificação, clusterização e regressão, onde sua utilização é justificada principalmente pela rapidez de treinamento e o custo custo computacional mais baixo comparado com redes neurais profundas ou métodos baseados em retropropagação. Wang et al. (Veja [15] e [16]) apresentam duas propostas do uso do ELM em problemas de Análise de Sobrevivência que consideram o modelo de Cox e o Modelo Buckley-James, e Kong et al. [17] utiliza um algoritmo Boosting L2 baseado em ELM e o modelo Buckley-James, mostrando uma melhora nas previsões de tempos de sobrevivência quando os modelos são não lineares.

O modelo de Cox de Riscos Proporcionais [18], é amplamente utilizado em Análise de Sobrevivência e, apesar de sua complexidade computacional ser baixa, depende da suposição de riscos proporcionais (a razão de riscos entre diferentes grupos ou indivíduos, controlada pelas covariáveis, é constante ao longo do tempo), a qual é difícil de ser satisfeita em problemas reais. A interpretação deste modelo é feita a partir da razão de riscos e o tempo de sobrevivência é obtido indiretamente. Em contrapartida, o modelo de Buckley-James (Buckley, 1979 [19]) estima diretamente o tempo de sobrevivência e dispensa a suposição de riscos proporcionais, apresentando o modelo como uma adaptação da regressão linear que lida com censura, usando mínimos quadrados e substituindo os valores censurados pelas suas esperanças condicionais.

A abordagem tradicional em modelos de regressão pressupõe, em geral, a normalidade dos erros e emprega o método de mínimos quadrados na estimação dos parâmetros. No entanto, um modelo normal sofre de falta de robustez no sentido de ser muito sensível quando existem observações extremas (*outliers*). Assim, relaxar a suposição de normalidade, utilizando uma distribuição simétrica com caudas mais pesadas que a distribuição normal, tem-se mostrado uma alternativa interessante para reduzir a influência dos dados extremos no processo de estimação. Muitos trabalhos sobre a utilização destas distribuições simétricas foram apresentados ao longo das últimas duas décadas, dentre eles, Massuia et al [20] apresenta um modelo de regressão *t-Student* para dados censurados. Neste contexto, a ideia de considerar um modelo Buckley-James com distribuição simétrica diferente da normal (especificamente *t-Student*) torna-se possível.

Considerando todos os elementos apresentados, o algoritmo de Aprendizado de Máquinas proposto considera o modelo Buckley-James robusto com distribuição de erros *t-Student* para o melhor tratamento de dados extremos, assim como a inclusão do ELM e

o Boosting L2 para o tratamento de não linearidade.

1.3 OBJETIVOS

O objetivo principal deste trabalho de doutorado é propor um algoritmo para o modelo de sobrevivência semiparamétrico que combina a ideia do modelo Buckley-James [19] considerando uma distribuição de probabilidade *t-Student*, robusta à presença de dados extremos (*outliers*) em lugar da distribuição normal com o uso da Máquina de Aprendizado Extremo (ELM) para a busca da melhor solução na predição de dados de sobrevida, com posterior aplicação e validação do algoritmo proposto na análise de dados de pacientes com doença renal crônica em tratamento por diálise peritoneal.

Os objetivos específicos do trabalho podem ser resumidos em:

(i) Apresentar detalhadamente o algoritmo proposto descrevendo as suposições e os parâmetros que devem ser considerados.

(ii) Avaliar os resultados da aplicação do algoritmo proposto a partir de diversos estudos de simulação de Monte Carlo. Os estudos de simulação incluem avaliações do algoritmo sob diversas caraterísticas de conjunto de dados e comparações com outras metodologias apresentadas na literatura.

(iii) Aplicar o algoritmo proposto em dados reais conhecidos na literatura e comparar os resultados com as outras metodologias apresentadas nos estudos de simulação.

(iv) Analisar os dados médicos de pacientes associados ao tratamento de Doença Renal Crônica (DRC), aplicando o algoritmo proposto e as outras metodologias apresentadas neste trabalho.

Dentre as metodologias disponíveis na literatura para Análise de Sobrevivência, este trabalho concentra suas comparações principalmente no modelo de Cox, por ser o mais difundido e amplamente utilizado nesse tipo de estudo. Além disso, são considerados modelos baseados na formulação de Buckley–James, descritos em estudos prévios, de modo a avaliar o desempenho da proposta em relação a metodologias clássicas e semiparamétricas consolidadas.

1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho de tese contêm seis capítulos, sendo organizados como segue abaixo.

Uma revisão bibliográfica sobre os principais modelos utilizados na Análise de Sobrevivência é apresentada no Capítulo 2, com ênfase tanto em métodos estatísticos clássicos quanto em abordagens baseadas em Aprendizado de Máquina. Os métodos estatísticos são agrupados em três grandes categorias: não paramétricos, como o estimador

de Kaplan-Meier; semiparamétricos, com destaque para o modelo de riscos proporcionais de Cox e o modelo de Buckley–James, adequado para lidar com censura; e paramétricos, que incluem modelos baseados em distribuições conhecidas, como exponencial, Weibull e log-normal. No campo do Aprendizado de Máquina, são discutidas técnicas adaptadas para dados censurados, com destaque para as árvores de sobrevivência (como as Random Survival Forests), as redes neurais profundas (como DeepSurv e DeepHit) e as máquinas de vetores de suporte para sobrevivência (Survival SVM). Tais abordagens vêm ganhando espaço por sua capacidade de modelar relações complexas entre covariáveis e tempos de evento, especialmente em contextos com grandes volumes de dados e múltiplas fontes de informação.

No Capítulo 3, apresenta-se a fundamentação teórica necessária e, na sequência, a metodologia desenvolvida neste trabalho. Os conceitos associados à Análise de Sobrevivência com ênfase em censura a direita; o modelo de Buckley-James; a técnica de Máquina de Aprendizado Extremo (ELM), a distribuição de probabilidade *t-Student* como alternativa à normal para o tratamento robusto de dados extremos (*outliers*); o modelo de Boosting; e o teste RESET de Ramsey, utilizado para avaliar a presença de não linearidade. Finalmente, é apresentada a forma como esses fundamentos são integrados, culminando na proposta do algoritmo desenvolvido para o tratamento de dados de sobrevivência e previsão de tempos de vida.

Os resultados de diversos estudos de simulação para avaliação do algoritmo proposto, bem como sua aplicação em dados reais conhecidos na literatura, são apresentados no Capítulo 4. A comparação com outras técnicas conhecidas em Análise de Sobrevivência é também discutida. As medidas de desempenho C-index e IBS, utilizadas nas avaliações quantitativas, são descritas e definidas neste capitulo.

O Capítulo 5 apresenta a análise detalhada dos dados sobre doença renal crônica em pacientes submetidos à diálise peritoneal, incluindo a descrição do estudo, a construção do banco de dados para análise, as variáveis utilizadas e uma caracterização inicial por meio de análises descritivas. Em seguida, são apresentados os resultados obtidos, que abrangem a aplicação do algoritmo proposto e sua comparação com as demais metodologias consideradas neste trabalho.

Finalmente, o Capítulo 6 apresenta as conclusões e considerações finais deste trabalho, reunindo os principais resultados obtidos e discutindo suas implicações no contexto da Análise de Sobrevivência. O capítulo também destaca as limitações observadas e propõe possíveis direções para pesquisas futuras.

2 REVISÃO BIBLIOGRÁFICA

As técnicas propostas para a Análise de Sobrevivência buscam modelar o tempo em que ocorre um evento de interesse considerando a presença de censuras, tendo como objetivo principal realizar previsões do tempo de sobrevivência e estimar a probabilidade de sobrevivência no tempo de sobrevivência estimado. Para este fim, nos primeiros desenvolvimentos da área foram apresentados alguns métodos estatísticos para o tratamento de dados com censura que estão fundamentadas em teorias probabilísticas. Em pesquisas recentes, diversos algoritmos de Aprendizado de Máquina foram adaptados para lidar com dados censurados num contexto onde os dados reais apresentam algumas caraterísticas desafiadoras que, em muitos casos, levam à violação de algumas suposições que os modelos estatísticos tradicionais, resultando em desempenho reduzido ou limitações práticas. A seguir apresenta-se uma visão global sobre os métodos de Análise de Sobrevivência mais conhecidos na literatura.

2.1 MÉTODOS ESTATÍSTICOS

Os métodos estatísticos tradicionais de Análise de Sobrevivência foram desenvolvidos com o objetivo de caracterizar probabilisticamente os tempos até a ocorrência de um evento de interesse, definindo as propriedades estatísticas dos estimadores dos parâmetros do modelo e da curva de sobrevivência. Para esse fim, diferentes abordagens foram propostas, cada uma baseada em suposições específicas sobre a distribuição dos tempos de sobrevivência e sobre a natureza das variáveis aleatórias envolvidas. Quando essas suposições são satisfeitas, tais métodos oferecem estimativas consistentes e interpretáveis, especialmente em cenários de baixa dimensionalidade. No entanto, em aplicações reais, muitas dessas condições não são plenamente atendidas, o que motivou o desenvolvimento de novas metodologias ou adaptações das existentes. Entre essas extensões modernas destacam-se os métodos de regularização, como o Lasso, que permitem lidar com dados de alta dimensionalidade e complexidade, comuns em estudos genômicos e biomédicos.

Uma classificação amplamente utilizada para organizar esses métodos está associada às suposições sobre a distribuição dos dados e a forma das relações entre variáveis. Assim, eles podem ser agrupados em modelos paramétricos, semiparamétricos e não paramétricos. Essa forma de apresentação pode ser encontrada em Colosimo et al. [42] e Wang et al. [7], entre outros, e constitui a base para a compreensão das diferentes estratégias de modelagem em sobrevivência.

Os modelos não paramétricos não fazem suposições fortes sobre a distribuição dos dados tornando-se mais flexíveis ao modelar relações complexas, porém podem apresentar menos eficiência nas estimações (especialmente com amostras pequenas) e maior dificuldade nas interpretações. Estes modelos são úteis quando não se conhece a distribuição explícita

dos dados e alcançam uma boa precisão com amostras grandes.

Os métodos não paramétricos são mais eficientes quando não há distribuição subjacente para o tempo de evento ou a suposição de risco proporcional não se mantém, o objetivo é a obtenção de uma estimativa empírica da função de sobrevivência. O método não paramétrico mais conhecido e usado é o de Kaplan-Meier (KM) proposto em 1958 ([21]). De forma geral, qualquer estimador de KM para a probabilidade de sobrevivência no tempo de sobrevivência especificado será um produto da mesma estimativa até o tempo anterior e a taxa de sobrevivência observada para esse tempo determinado, por isto, o método KM também é referido como um método de limite de produto. Um segundo método é o de Nelson-Aalen (NA) ([22] e [23]), que utiliza na sua construção algumas técnicas modernas de processos de contagem, este apresenta essencialmente as mesmas caraterística de KM. Finalmente, método de Tabela de Vida (LT) (Cutler et al. 1958 [24]) é construído a partir da aplicação do método KM a dados de sobrevivência agrupados por intervalos.

Os modelos semiparamétricos combinam componentes paramétricos e não paramétricos, fazendo algumas suposições paramétricas e flexibilizando certas suposições no conjunto de dados, são interpretáveis com dificuldade mas com uma implementação mais complexa e com menor eficiência que os modelos paramétricos pois as distribuições dos resultados não são conhecidas.

O modelo de Cox ([18] e [25]), também conhecido como modelo de riscos proporcionais de Cox, é um dos métodos mais utilizados na Análise de Sobrevivência. Ele é um modelo semiparamétrico, o que significa que ele faz algumas suposições paramétricas sobre os efeitos das covariáveis, mas não assume uma distribuição específica para o tempo de sobrevivência. No modelo de Cox, a taxa de risco (ou função de risco) para um indivíduo em um determinado tempo, dado um conjunto de covariáveis, é modelada considerando uma taxa de risco de base (a parte não paramétrica do modelo) e outra parte paramétrica, que descreve como as covariáveis influenciam o risco. O modelo de Cox é amplamente utilizado quando o interesse está em avaliar o efeito das covariáveis sobre o risco de um evento, como morte ou falência de um equipamento, sem a necessidade de especificar a forma exata da função de risco ao longo do tempo.

As principais suposições do modelo de Cox incluem a de riscos proporcionais, o que implica que a razão de risco entre dois indivíduos é constante ao longo do tempo, e que as covariáveis têm um efeito multiplicativo sobre o risco. Essa suposição pode ser limitante em alguns casos, como quando os efeitos das covariáveis mudam com o tempo (produzindo uma violação da suposição de proporcionalidade). As vantagens do modelo de Cox incluem sua flexibilidade, a possibilidade de incluir múltiplas covariáveis, quando existem dados censurados. No entanto, uma das desvantagens é que, se a suposição de riscos proporcionais for violada, as inferências podem ser equivocadas, este modelo será tratado com mais detalhe no próximo capítulo.

Existem várias variações do modelo de Cox que são úteis em contextos específicos. O modelo de Cox estratificado permite que a taxa de risco de base varie entre estratos (grupos) sem assumir uma estrutura paramétrica para o efeito de grupo (Veja Klein, 2003 [26] e Therneau, 2000 [27]) . O modelo de Cox com covariáveis dependentes do tempo permite que as covariáveis mudem ao longo do tempo, acomodando a possibilidade de variações nos efeitos das covariáveis sobre o risco (Kalbfleisch, 2002 [28]). Outra variação é o modelo de Cox penalizado, que é útil para seleção de variáveis e regularização em problemas de alta dimensionalidade, sendo frequentemente empregado em Aprendizado de Máquina, onde o CoxNet (uma versão regularizada com penalidade LASSO ou Ridge) pode ser usado para seleção de variáveis em grandes conjuntos de dados (Veja Tibshirani, 1997 [31], Simon et al., 2011 [31] e Hastie et al.,2009 [33]).

O modelo de Buckley–James (BJ), apresentado em 1979 [19], é uma abordagem semiparamétrica alternativa desenvolvida para lidar com dados de sobrevivência. Ele pode ser considerado uma extensão do modelo de regressão linear para dados censurados, adaptado ao contexto de sobrevivência, permitindo incorporar observações censuradas na estimação dos parâmetros de regressão. Diferentemente do modelo de Cox, o BJ não assume a proporcionalidade de riscos, uma condição que, em situações reais, frequentemente não é satisfeita, o que o torna uma alternativa relevante para modelagem de sobrevivência. Este modelo estima os tempos de sobrevivência diretamente, considerando o efeito das covariáveis. A estimação dos tempos censurados é realizada com base na função de sobrevivência obtida pelo método de Kaplan–Meier (KM), e em seguida ajusta-se um modelo linear para o logaritmo do tempo de sobrevivência, considerando simultaneamente as observações não censuradas e as aproximações dos tempos censurados.

O modelo de BJ é particularmente útil quando se deseja evitar as suposições fortes dos modelos paramétricos (como a distribuição exata dos tempos de sobrevivência), mas ainda assim obter uma estimativa confiável do efeito das covariáveis sobre o tempo de sobrevivência. Se apresenta como uma boa alternativa ao modelo de Cox quando a suposição de riscos proporcionais não é válida. Para lidar com dados de sobrevivência de alta dimensionalidade, Wang et al. (2008) [34] aplicaram o regularizador Elastic Net na regressão BJ.

Os métodos paramétricos (Veja por exemplo Colosimo et al[42]) assumem distribuições específicas para os dados e são baseados em um número fixo de parâmetros que são determinados pela forma funcional do modelo proposto. As principais suposições necessárias para este tipo de abordagem incluem: a especificação de uma distribuição conhecida para os tempos de sobrevivência e a independência estatística entre as observações, o que implica, consequentemente, uma distribuição probabilística para os erros. Quando as suposições são cumpridas, estes modelos se tornam mais eficientes e fáceis de interpretar. No entanto, em situações reais, muitas das suposições podem não ser cumpridas e ao ter uma forma funcional fixa podem falhar se existem relações mais complexas nos dados o

que os torna pouco ótimos.

Este tipo de modelos constitui uma alternativa aos modelos semiparamétricos, pois assume-se que os tempos de sobrevivência seguem uma distribuição de probabilidade previamente especificada. Além da distribuição normal (assumida para o logaritmo do tempo, como no caso do modelo de Buckley–James), outras distribuições frequentemente utilizadas incluem a exponencial, Weibull, logística, log-normal e log-logística, todas definidas para valores positivos, como ocorre com o tempo de sobrevivência. Os modelos de regressão paramétrica censurada partem da hipótese de que os tempos de sobrevivência de todas as instâncias seguem uma distribuição teórica particular (Lee et al., 2003 [35]). Esses modelos oferecem alternativas relevantes aos semiparamétricos baseados em Cox e são amplamente aplicados em diversos domínios, fornecendo uma abordagem simples e eficiente para prever o tempo até o evento de interesse. Em geral, modelos paramétricos de sobrevivência produzem estimativas consistentes com a distribuição teórica assumida, o que pode ser vantajoso quando essa suposição é plausível para os dados analisados.

O Modelo de Tempo de Falha Acelerado (AFT) é caracterizado por uma relação linear entre o logaritmo do tempo de sobrevivência e as covariáveis. O termo de erro segue uma distribuição semelhante ao do logaritmo do tempo de sobrevivência. Normalmente, considera-se de forma paramétrica que essa variável de erro segue uma das distribuições mencionadas no parágrafo anterior. Nesse caso, a sobrevivência depende tanto da covariável quanto da distribuição subjacente. Isso significa que a única distinção entre um modelo AFT e os métodos lineares regulares seria a inclusão das informações censuradas no problema de análise de sobrevivência. O modelo AFT é aditivo com relação ao logaritmo do tempo de sobrevivência mas multiplicativo com relação a tempo de sobrevivência. Kalbfleisch et al. [28] é uma excelente fonte de referência sobre o modelo AFT e outros modelos paramétricos de sobrevivência.

2.2 MÉTODOS DE APRENDIZADO DE MÁQUINA

Os métodos de Aprendizado de Máquina são propostos com o intuito de prever a ocorrência de eventos em um determinado momento, isto, é determinar o tempo de sobrevida. Os algoritmos propostos combinam métodos estatísticos conhecidos e métodos de Aprendizado de Máquinas. Esta combinação os torna mais eficientes quando se tem dados de alta dimensionalidade e com maior flexibilidade para tratar as dependências e não linearidades das covariáveis e os diversos comportamentos dos tempos de sobrevivência.

Dentre as várias abordagens básicas de Aprendizado de Máquina desenvolvidas para a Análise de Sobrevivência descrevemos as mais conhecidas: As árvores de sobrevivência, as redes neurais e máquinas de vetores suporte.

As árvores de sobrevivência (Veja, Bou-Hamad et al. [30]) são uma extensão das árvores de decisão (Breiman et al. [29]), utilizadas para analisar dados de sobrevivência.

Assim como as árvores de decisão para classificação e regressão, essas árvores segmentam os dados em subgrupos com base nas covariáveis, mas, no caso de sobrevivência, o objetivo é modelar o tempo até o evento de interesse (como falha ou morte) e lidar com dados censurados (quando o evento não ocorre para todos os indivíduos durante o período de estudo).

No contexto de árvores de sobrevivência, a divisão dos dados é feita com base na maximização da diferença nas funções de sobrevivência (ou risco) entre os grupos resultantes de cada divisão. Em vez de prever uma classe ou valor numérico, a árvore de sobrevivência estima a função de sobrevivência ou a função de risco para cada nó terminal (grupo), o que permite comparar como diferentes covariáveis influenciam o tempo de sobrevivência.

Entre as vantagens deste método pode-se mencionar a flexibilidade, pois as árvores de sobrevivência não fazem suposições paramétricas sobre a forma da função de risco ou da função de sobrevivência, tornando-as uma abordagem não paramétrica poderosa. Também quanto à facilidade de interpretação, pois as árvores de sobrevivência criam uma estrutura hierárquica de divisões com base em variáveis explicativas fáceis de visualizar. Finalmente, pode-se capturar interações complexas entre covariáveis sem a necessidade de especificá-las explicitamente. Por outro lado, as árvores de sobrevivência podem ser sensíveis a pequenas variações nos dados, levando a diferentes divisões, o que pode impactar a estabilidade do modelo.

Uma variação popular das árvores de sobrevivência é o método Random Survival Forests (RSF) (veja Ishwaran et al., 2007 [36], que combina várias árvores de sobrevivência para aumentar a precisão e a estabilidade do modelo, aplicando uma técnica de agregação para gerar várias árvores a partir de amostras diferentes dos dados.

As redes neurais têm sido aplicadas à Análise de Sobrevivência para modelar dados complexos e não lineares, especialmente em cenários onde métodos tradicionais, como o modelo de Cox, podem não capturar adequadamente as relações entre covariáveis e o tempo de sobrevivência. Um dos primeiros trabalhos importantes nessa área foi o de Faraggi e Simon (1995) [37], que introduziu uma rede neural baseada em máxima verossimilhança para dados censurados, inspirada no modelo de riscos proporcionais de Cox. A rede neural utiliza a função de verossimilhança parcial do modelo de Cox como função de perda, permitindo à rede aprender a função de risco proporcional de forma mais flexível e não linear. Diferente do modelo de Cox clássico, que pressupõe uma relação linear entre as covariáveis e o logaritmo do risco, a rede neural proposta por eles pode capturar relações não lineares entre as covariáveis e o tempo de sobrevivência. Ao maximizar a verossimilhança durante o treinamento, a rede neural é capaz de prever o tempo de sobrevivência e lidar com dados censurados de maneira eficaz.

Apesar da rede neural oferecer uma generalização não linear do modelo de Cox,

permitindo modelar dados complexos que envolvem interações complicadas entre as variáveis preditoras existe um problema da interpretação. Enquanto o modelo de Cox oferece coeficientes que podem ser interpretados diretamente, as redes neurais, com suas múltiplas camadas e nós, dificultam a interpretação dos efeitos individuais das covariáveis. Além disso, o risco de sobreajuste é mais elevado em redes neurais, especialmente quando há poucos dados disponíveis.

Uma expansão da proposta de Faraggi, denominada DeepSurv, foi desenvolvida por Katzman et al. em 2018 [38]). Este modelo utiliza redes neurais profundas para generalizar o modelo de Cox em contextos mais modernos e computacionalmente avançados. Essas redes podem lidar com dados de alta dimensionalidade, oferecendo previsões mais precisas ao capturar interações não lineares complexas entre covariáveis.

O ELM (Huang et al, 2006 [13]) é uma técnica de redes neurais que também foi adaptada para a Análise de Sobrevivência, oferecendo uma abordagem rápida e eficiente para modelar dados complexos com censura. Por ser uma rede neural de camada única destaca-se pela sua capacidade de treinamento extremamente rápido, uma vez que os pesos das camadas ocultas são gerados aleatoriamente e os parâmetros de saída são ajustados através de uma simples inversão de matriz. A adaptação de ELM para sobrevivência envolve ajustar a função de perda para incorporar dados censurados, semelhante ao que é feito com redes neurais tradicionais, mas preservando a rapidez de treinamento que é característica do ELM.

Uma abordagem comum é combinar o modelo de riscos proporcionais de Cox ou outro modelo conhecido com a arquitetura ELM, permitindo que a rede aprenda uma relação entre as covariáveis e o tempo de sobrevivência sem precisar assumir linearidade ou relações paramétricas rígidas.

Conforme apresentado por Huang et al. [12, 13], a alta velocidade de treinamento em comparação às redes neurais tradicionais oferece vantagem quando é preciso analisar grandes conjuntos de dados, como em estudos de coorte médica ou dados genômicos. Por outro lado, a arquitetura do Extreme Learning Machine (ELM) é simples, com poucos parâmetros a serem ajustados, o que pode reduzir a suscetibilidade a problemas de ajuste excessivo em relação a arquiteturas mais complexas. Assim como outras redes neurais, o ELM mantém a capacidade de capturar relações não lineares entre covariáveis e o tempo de sobrevivência.

As Máquinas de Vetores Suporte (SVM) foram adaptadas para lidar com dados censurados, como tempos de sobrevivência ou falha, oferecendo uma alternativa não paramétrica ao modelo de Cox e outras abordagens tradicionais. As SVMs convencionais são projetadas para resolver problemas de classificação e regressão com margens máximas, mas quando aplicadas à análise de sobrevivência, a principal dificuldade está em lidar com dados censurados — casos em que o evento de interesse (como morte ou falha)

não foi observado dentro do período de estudo. Para adaptar as SVMs a este contexto, surgiram diferentes variações, como as Rank-SVM ou Survival-SVM, que incorporam informações de censura no processo de modelagem. Van Belle et al. (2011) [39], oferece uma comparação detalhada entre diferentes métodos baseados em SVM para análise de sobrevivência, incluindo o Survival-SVM e o Rank-SVM, discutindo suas respectivas vantagens e desvantagens em cenários de dados censurados.

A Rank-SVM para sobrevivência é uma adaptação das Máquinas de Vetores Suporte (SVM) focada na ordenação (ranking) dos tempos de sobrevivência, em vez de prever diretamente o tempo de falha. Essa técnica é útil para dados censurados, onde nem todos os eventos de interesse são observados. A Rank-SVM ordena os tempos de sobrevivência com base no risco, permitindo que a censura seja tratada de maneira eficiente. A ideia principal é formular o problema como um aprendizado por ranking, onde a função objetivo visa a maximização da margem entre pares de observações ordenadas.

O Survival-SVM é uma extensão das Máquinas de Vetores Suporte (SVM) que lida especificamente com dados de sobrevivência e dados censurados. Essa abordagem combina a flexibilidade das SVMs com as necessidades específicas da análise de sobrevivência, permitindo a modelagem de funções de risco ou diretamente dos tempos de sobrevivência. Diferente do Rank-SVM, que foca em ordenar tempos de sobrevivência, o Survival-SVM busca prever o risco de falha ou o tempo de sobrevivência de maneira mais direta.

Van Belle et al. 2011 [39] apresenta uma comparação detalhada entre diferentes métodos baseados em SVM para análise de sobrevivência, incluindo o Survival-SVM e o Rank-SVM, discutindo suas respectivas vantagens e desvantagens em cenários de dados censurados.

Finalmente, métodos de Boosting são especialmente úteis em problemas com alta dimensionalidade ou quando existem muitas covariáveis que podem influenciar o risco de falha, como em dados genômicos, medicina personalizada e em grandes estudos clínicos. Como o Boosting trabalha iterativamente corrigindo erros, ele pode lidar bem com ruído nos dados, oferecendo modelos preditivos robustos. Um dos métodos mais comuns de Boosting em análise de sobrevivência é o CoxBoost, que adapta o modelo de riscos proporcionais de Cox à estrutura do Boosting. Ele aplica o conceito de Boosting para selecionar covariáveis e construir um modelo forte com base na função de risco do modelo de Cox. Neste contexto, entende-se por modelo forte a combinação de vários modelos fracos — cada um com desempenho limitado isoladamente — em um preditor único e mais robusto, capaz de alcançar maior precisão e poder explicativo na análise de sobrevivência. Outro exemplo é o Gradient Boosting Machine (GBM) adaptado para dados de sobrevivência, que usa a função de perda do log-partial likelihood do modelo de Cox na orientação do aprendizado.

3 FUNDAMENTAÇÃO E DESENVOLVIMENTO METODOLÓGICO

Neste capítulo, são apresentados os principais conceitos teóricos que fundamentam a construção do algoritmo de aprendizado de máquinas que será desenvolvido neste trabalho. Primeiramente, são introduzidos os conceitos teóricos da Análise de Sobrevivência, onde descreve-se com detalhe o estimador de Kaplan-Meier e Modelo de Riscos Proporcionais de Cox devido a serem muito utilizados nas técnicas de Aprendizado de Máquinas para dados de sobrevivência. Um exemplo simples é apresentado para um melhor entendimento dos conceitos e modelos descritos. A seguir, apresentam-se os modelos BJ e um modelo de regressão linear com erros com distribuição *t-Student*. Duas técnicas de aprendizado de máquinas, o ELM e o algoritmo boosting L2 são apresentados para concluir as ferramentas que serão usadas no trabalho. Finalmente, descreve-se como esses fundamentos teóricos são integrados na formulação metodológica, culminando na proposta do algoritmo para o tratamento de dados de sobrevivência e previsão de tempos de vida.

3.1 ANÁLISE DE SOBREVIVÊNCIA

Na análise de sobrevivência, a variável resposta corresponde ao tempo transcorrido até a ocorrência de um evento de interesse. É usual na literatura defini-la como *tempo de falha*. Apesar da conotação negativa da frase, ela não necessariamente é uma situação desfavorável, um exemplo disto é o *tempo de reinserção ao mercado de trabalho*.

3.1.1 **Tipos de estudos**

O planejamento do procedimento para a obtenção de dados é importante para a Análise de Sobrevivência. Estes são obtidos a partir de estudos longitudinais, isto é, os registros são realizados ao longo de um intervalo de tempo predefinido, portanto, com custos mais altos que os estudos transversais e adicionalmente, investe-se um período de tempo longo na coleta de informações. Por isto, a necessidade de realizar um estudo planejado e sistematizado que otimize custos de tempo e dinheiro torna-se primordial.

Na área da saúde, alguns tipos de estudos clínicos são mais conhecidos, algumas caraterísticas específicas os diferenciam. Quando se planeja um estudo observacional e prospectivo podem ser realizados dois tipos: o descritivo, no qual se acompanha uma amostra de doentes e identifica-se alguns fatores de risco para a doença, e o estudo denominado coorte onde dois grupos que foram expostos ou não a um fator de interesse são acompanhados num período de tempo para avaliar a incidência ou doença de interesse. Um estudo observacional retrospectivo é denominado caso-controle, no qual também se comparam dois grupos (por exemplo, doentes e não doentes) avaliando diversos fatores de interesse, neste estudo utilizam-se informações já conhecidas como o histórico clínico, assim estes resultam de custos mais baratos, porém podem ser menos precisos, pois dependem da

qualidade da informação histórica obtida. Finalmente, tem-se o estudo clínico aleatorizado que é prospectivo e experimental, no qual aloca-se de forma aleatória os tratamentos aos pacientes. O importante nestes estudos é a definição e a obtenção do tempo até a ocorrência do evento de interesse, a literatura mostra que aplicações da análise de sobrevivência são mais frequentes sobre estudos de coorte e clínicos aleatorizados. Informações mais detalhadas sobre estes estudos pode ser encontrado em Rothman et al., (1998) [40].

Estudos na área de confiabilidade são geralmente experimentais, planejados simulando situações reais e são realizados em áreas de testes nas empresas. Um caso interessante são os teste de vida acelerados (Veja Nelson et al, 1990 [41]), no qual as unidades amostrais são estressadas com a finalidade de que falhem mais rápido reduzindo o tempo do experimento, o que se torna mais útil em testes de item que por sua natureza tem durabilidade grande.

Estudos epidemiológicos, denominados coortes, frequentemente são conduzidos ao longo de anos e com custos altos, portanto um cuidado no planejamento, sistematização e acompanhamento do estudo é muito importante. O mesmo deve ser aplicado em estudos de confiabilidade, onde se acompanha o tempo de durabilidade de uma peça, por exemplo, devendo ser entendido completamente para determinar algumas caraterísticas do mesmo.

3.1.2 **Modelagem de dados de sobrevivência**

Um modelo estatístico tem o objetivo de explicar as relações existentes num conjunto de dados considerando uma relação funcional que descreva da melhor maneira possível o padrão observado, seja para fins de ajuste, previsão ou interpretação. Para que essa aproximação seja considerada estatisticamente válida, é necessário que certas suposições sejam atendidas, como a independência das observações, a especificação correta da forma funcional do modelo e, em alguns casos, a definição de uma distribuição probabilística para os erros ou para os tempos de sobrevivência. Neste contexto, modelos de sobrevivência apresentam algumas caraterísticas específicas que devem ser consideradas durante as análises. Estes modelos vem mostrando um crescimento desde o final do século passado. Nos tempos atuais, a intensificação e aperfeiçoamento das técnicas estatísticas junto ao desenvolvimento acelerado das ferramentas computacionais, especialmente dos algoritmos de aprendizado de máquinas, explicam o seu grande avanço e uso frequente.

Os métodos estatísticos desenvolvidos e mais usados incluem os não paramétricos no qual a estimação de Kaplan-Meier é mais utilizada, os semiparamétricos, onde o modelos de Cox são os mais conhecidos, e finalmente os modelos paramétricos que utilizam distribuições de probabilidade conhecidas como a exponencial, weibull ou logística para modelar a variável de interesse, neste caso, o tempo. Por outro lado, os métodos de Aprendizado de Máquinas incluem as árvores de sobrevivência, redes neurais, máquinas de vetores suporte e comitê de classificadores. Uma descrição dos diversos métodos usados na

análise de sobrevivência podem ser encontrada em [7].

As aplicações da Análise de Sobrevivência podem ser encontradas em diversas áreas de pesquisa, estas são referenciadas em [42] e [7]. As mais frequentes ocorrem na área médica, onde se estuda o tempo de ocorrência de um evento associado a uma doença (uma situação comum em estudos de câncer é a recidiva, isto é, o retorno da doença após tratamento, em vez do registro de cura ou morte). Em tempos recentes, a inclusão de informações de expressões genéticas levam ao tratamento de dados com alta dimensionalidade. Assim, desenvolvimentos de novas técnicas para análises de sobrevivência que contornem esta situação foram apresentados, veja por exemplo [32] e [43].

Na área de engenharia, problemas de confiabilidade de dispositivos ou sistemas usam a Análise de Sobrevivência para avaliar o tempo de falha de alguns componentes. Nas áreas de economia e sociologia utiliza-se para analisar históricos de eventos como nascimentos, casamentos, acessos a empregos, transações em mercados financeiros, retenção de estudantes entre outros. Finalmente, o crescimento do uso da internet, levou ao uso da análise de sobrevivência para avaliar os tempos para retorno sobre financiamentos coletivos, para compra de determinado serviço, para acesso a propaganda e outros. Desta forma, pode-se considerar que é uma técnica com uma perspectiva promissora em aplicações futuras.

3.1.3 **Caraterização dos dados de sobrevivência**

A seguir, descrevem-se formalmente algumas características fundamentais dos dados utilizados em análises de sobrevivência, tomando como referência os trabalhos de Colosimo et al. [42] e Wang et al. [7].

3.1.3.1 *Obtenção de dados*

Os dados para as análises são obtidos ao longo de uma dimensão que apresente um ordenamento, em geral considera-se o tempo. Em estudos médicos este tipo de coleta de dados é denominado coorte, já no caso de estudos sociais estes são conhecidos como painéis. São planejados para acontecer num período longo de tempo, porém devem ter uma data determinada para sua finalização.

3.1.3.2 *Variável resposta*

A variável resposta corresponde ao tempo de ocorrência do evento de interesse, é comumente denominada Tempo de Falha. Como mencionado anteriormente, o tempo é a medida mais comum para a variável resposta, no entanto, em problemas associados a confiabilidade algumas outras escalas de medida ordenadas são consideradas, por exemplo distância percorrida.

3.1.3.3 *Censura nos dados*

Mesmo considerando estudos que sejam realizados em períodos longos de tempo, tem-se uma data fixada para a finalização deste. Assim, espera-se que nem todos os indivíduos que formam parte do estudo falhem no período de tempo considerado, isto é, sobrevivem ao tempo determinado para sua finalização. Assim, esta caraterística, denominada censura, deve ser considerada. Esta caraterística à diferencia dos modelos de regressão linear clássica que são amplamente conhecidos.

Em análises de sobrevivência, a censura corresponde a uma observação incompleta da variável resposta. Ela pode ocorrer de três formas principais: à direita, quando o evento não é observado até o fim do acompanhamento; à esquerda, quando se sabe que o evento ocorreu antes de certo momento, mas não se conhece o tempo exato; e intervalar, quando o evento é apenas conhecido por ter acontecido dentro de um intervalo de tempo.

A censura à direita é definida como uma observação parcial (ou incompleta) da resposta. Isto é, quando um indivíduo não apresenta uma falha até o final estipulado do estudo, esta informação se torna incompleta, pois não é possível determinar um tempo exato de falha, porém é possível afirmar que o tempo de falha é maior que o tempo de finalização observada.

Formalmente, considerando uma censura aleatória à direita e definindo duas variáveis aleatórias T: Tempo de falha e C: Tempo de censura que é independente de T. No estudo observa-se:

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

Desta forma, considerando uma amostra de tamanho n, a representação dos dados de sobrevivência é o conjunto das n observações, onde a i-ésima observação, $i = 1, .., n$, é $(t_i,\delta_i,\mathbf{x}_i)$, onde t_i é o tempo de falha ou censura da observação i, δ_i é o indicador de falha ou censura na observação i adotando o valor 1 se é tempo de falha e 0 se é censura, e $\mathbf{x}_i$ o vetor de covariáveis medidas para a observação i.

A Figura 1 mostra um esquema de obtenção de dados com censura aleatória à direita na área médica que inclui duas variáveis explicativas (atributos): Idade e Grau de doença. O tempo de falha, nesta situação, corresponde ao óbito do paciente.

A seguir, descreve-se outro exemplo, desta vez relacionado à área de confiabilidade. Neste caso, o objetivo é modelar o tempo de falha de uma peça do motor de um veículo. Para isso, coletam-se dados sobre a peça do motor ao longo de um período de tempo predeterminado, registrando o tempo decorrido até a falha ocorrer. A presença de censura é considerada, pois assume-se que nem todos os veículos testados apresentarão falha na peça do motor até o final do período estipulado para o estudo. Algumas variáveis de interesse consideradas são a idade do veículo, quilometragem rodada e o desgaste das

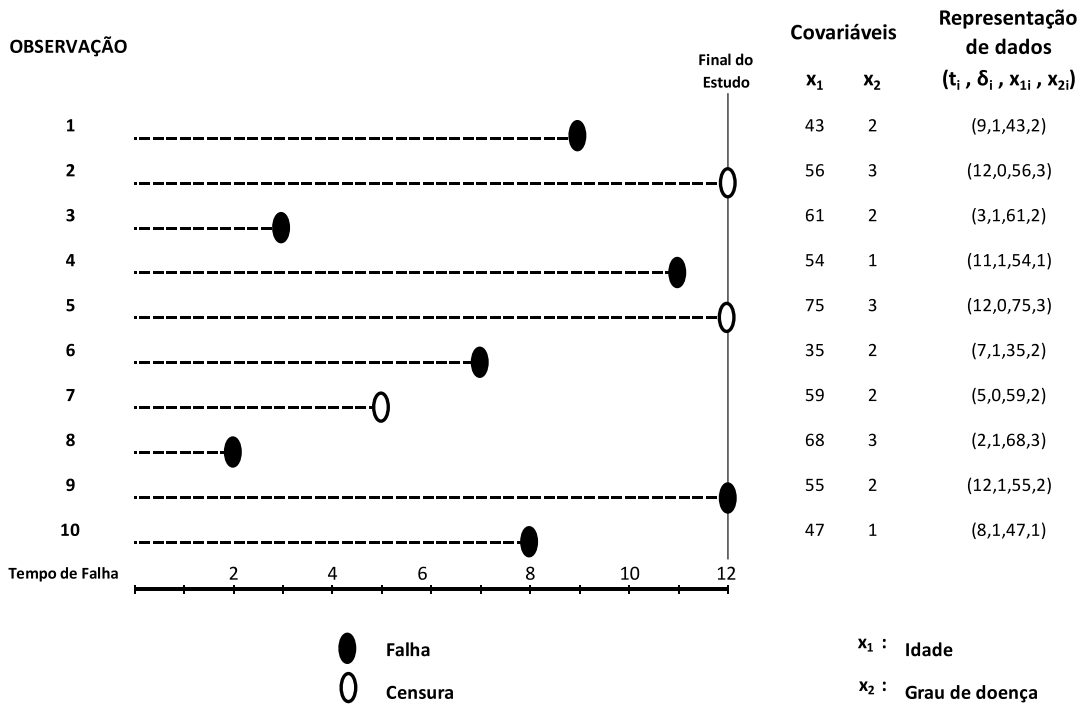

Figura 1 – Ilustração para conjunto hipotético de dados clínicos que apresentam censura aleatória.

peças do veículo.

A censura à direita é caracterizada quando o tempo de ocorrência de evento medido encontra-se ao lado direito do tempo registrado. Quando o tempo registrado é maior que o tempo de falha apresenta-se censura à esquerda. Em alguns casos pode-se apresentar simultaneamente censuras à direita e esquerda, isto é, duplamente censuradas. A metodologia utilizada para censura à direita pode ser utilizada nas outras situações desde que os dados sejam organizados adequadamente.

O fato de considerar os dados censurados na análise é justificado porque, mesmo parcialmente, esses dados fornecem informações valiosas sobre o tempo de falha. A eliminação dos dados censurados nas análises leva à obtenção de resultados enviesados, pois os tempos de falha que excedem os valores da censura não são considerados, embora, na realidade, eles ocorram.

3.1.4 O tempo de sobrevivência

O tempo de sobrevivência pode ser definida como uma variável aleatória que é usualmente contínua e que é especificada a partir de uma função de sobrevivência ou uma taxa de falha (ou risco).

A função de sobrevivência, denotada por $S(t)$, é definida como a probabilidade de

não falhar até um tempo t, isto é, sobreviver ao tempo t. Assim, tem-se que:

$$S(t) = P(T \geq t), \tag{3.1}$$

de onde a função de distribuição acumulada de T é $F_T(t) = 1 - S(t)$.

Para determinar a probabilidade de falha ocorrer no intervalo $[t_1, t_2)$ faz-se $S(t_1) - S(t_2)$. Finalmente, apresenta-se a taxa de falha no intervalo é a probabilidade ocorrer falha nesse intervalo, dado que não ocorreu antes de t_1 dividida pelo seu comprimento, dada por:

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

Define-se $\lambda(t)$ como a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t, obtida considerando $[t, \Delta t)$ com Δt pequeno. Assim, tem-se que:

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

A função de taxa de falha $\lambda(t)$ é maior que zero e descreve a distribuição de tempo de vida. Esta função é mais informativa que a função de sobrevivência, diferentes funções de sobrevivência podem ter formas semelhantes, porém suas funções de taxa de falha podem diferir drasticamente. Algumas relações importantes entre estas duas funções são:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{\partial}{\partial t}(log\, S(t))$$

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

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

3.2 ESTIMADOR DE KAPLAN-MEIER

Como sugerido por [44], o passo inicial na análise é resumir os tempos de sobrevivência t_i. O estimador de Kaplan-Meier [21] é utilizado para computar o tempo médio de sobrevivência considerando a presença de observações censuradas. O algoritmo a seguir descreve a obtenção deste estimador:

1. Determinar as k falhas distintas entre as n observações amostradas, identificando o tempo em que acontece cada uma delas ($k \leq n$).

2. Ordenar os k tempos identificados $t_1 < t_2 < ... < t_k$.

3. Calcular d_i, o número de observações que apresentam falha no tempo t_i, onde $i = 1, 2, ..., k$.

4. Calcular n_i, o número de observações sob risco no tempo t_i, onde $i = 1, 2, ..., k$. O valor corresponde às observações que não apresentam falha ou censura até o tempo t_i.

5. Calcular o estimador de Kaplan-Meier considerando a seguinte expressão

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right), \tag{3.2}$$

Observe que uma curva de sobrevivência Kaplan-Meier é calculada a partir de probabilidades que refletem o fato que para sobreviver a um tempo t deve-se superar todos os acontecimentos até o momento t.

3.3 MODELO DE RISCOS PROPORCIONAIS DE COX

O modelo de riscos proporcionais de Cox foi apresentado em 1972 ([18]) e desde então, é um dos artigos mais citados na área de estatística. Algumas características do modelo de Cox são descritas em [44] e [45] onde, as mais relevantes são:

- O modelo é intimamente ligado à curva de sobrevivência de Kaplan-Meier e estima as diferenças de risco experimentadas por grupos com diferentes características.
- O modelo de riscos proporcionais não depende de uma distribuição de probabilidade específica e assume que todos os grupos apresentam o mesmo risco basal (uma função do tempo) e que aumenta ou diminui de acordo a um fator multiplicativo que depende das características do grupo (covariáveis).
- O modelo proposto elimina o risco de linha basal na análise, utilizando para isto a verosimilhança parcial [25], assim é possível medir o efeito das covariáveis consideradas.
- O modelo lida adequadamente com censura (sobrevivência após o final do estudo), covariáveis e variações no tempo, situações comuns em problemas aplicados.

3.3.1 **Formalização do modelo**

O modelo de riscos proporcionais de Cox considera:

$$\lambda\left(t|\mathbf{x}\right) = \lambda_0\left(t\right) g\left(\mathbf{x}^\top \boldsymbol{\beta}\right), \tag{3.3}$$

onde os componentes do modelo são:
$\mathbf{x} = (x_1, x_2, ..., x_p)^\top$ são p covariáveis consideradas,
$\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^\top$ é o vetor de parâmetros,
$\lambda\left(t|\mathbf{x}\right)$ é a função de risco, isto é, a probabilidade de que um indivíduo com covariáveis $\mathbf{x}$

apresente falha no tempo t, dado que o indivíduo não falhou antes de t,
$g(.)$ é uma função não negativa, tal que $g(0) = 1$, e
$\lambda_0(t)$ é a função base, tal que $\lambda(t|\mathbf{x}) = \lambda_0(t)$ se $\mathbf{x} = \mathbf{0}$.

Este modelo é considerado semiparamétrico, assim é possível observar que a função de risco $\lambda(t|\mathbf{x})$ é explicada por dois componentes:

Uma componente não paramétrica, que corresponde à função base $\lambda_0(t)$, que é uma função não negativa do tempo t que não é especificada.

Uma componente paramétrica, que corresponde a função $g\left(\mathbf{x}^\top \boldsymbol{\beta}\right)$ que usualmente considera $exp\left\{\mathbf{x}^\top \boldsymbol{\beta}\right\} = exp\left\{\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p\right\}$. Observe que não se considera um parâmetro β_0 pois ele é absorvido por $\lambda_0(t)$.

A estimação dos parâmetros $\boldsymbol{\beta}$ é realizada a partir da verosimilhança parcial que usa a mesma construção proposta no método Kaplan-Meier de produtos no tempo, isto é, obtida ao calcular a probabilidade em cada evento de tempo t_k como um produto de probabilidades condicionais.

A função de verosimilhança parcial depende dos parâmetros $\boldsymbol{\beta}$, a qual é maximizada visando encontrar os estimadores de máximo verosimilhança parcial $\hat{\boldsymbol{\beta}}$. Adicionalmente, não é requerida a especificação da função base $\lambda_0(t)$, fazendo que o método seja flexível e robusto.

A suposição de riscos proporcionais implica que a relação entre a função de risco $\lambda(t|\mathbf{x})$ e a função base não depende do tempo, portanto é um fator constante que depende de $\mathbf{x}$ e dos parâmetros $\boldsymbol{\beta}$. Da mesma forma, ao considerar dois indivíduos diferentes i e j, onde $i \neq j$ e $i, j = 1, 2, ..., n$, a razão das funções de risco destes é constante e não dependem do tempo, isto é:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\, g\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{\lambda_0(t)\, g\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)} = exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_j^\top \boldsymbol{\beta}\right\} \tag{3.4}$$

O modelo de riscos proporcionais de Cox permite realizar inferências, isto é, obter estimativa de parâmetros, erros padrão e intervalos de confiança, da mesma forma que o modelo de regressão usual permite. Além disto, permite modelar covariáveis dependentes do tempo, isto é, covariáveis medidas várias vezes durante o estudo. O modelo é aplicável para todos os tipos de censura descritos. Adicionalmente, existem versões penalizadas que lidam com dados com alta dimensão e outras que utilizam ferramentas de aprendizado de máquinas como é o caso de florestas aleatórias e aprendizado profundo (Veja por exemplo [16], [32], [38], [43] entre outros).

3.4 UM EXEMPLO DE ANÁLISE DE SOBREVIVÊNCIA

Objetivo deste exemplo é utilizar o estimador Kaplan-Meier e o modelo de Cox num conjunto de dados clínicos reais para entender a relação entre estas duas técnicas e

entender alguns pontos específicos de cada uma delas.

O conjunto de dados analisado é apresentado em [42] e corresponde aos resultado de um estudo clínico aleatorizado em 29 pacientes que investigou o efeito de uma terapia com esteroide no tratamento de hepatite viral aguda. O estudo durou 16 semanas e registrou-se o tempo até a morte do paciente. A censura é registrada quando o acompanhamento é perdido ou o estudo é finalizado e o paciente continua vivo. Dos 29 pacientes, 15 formam parte do grupo controle e 14 receberam a terapia com esteroide . Os dados são apresentados na Tabela 1.

Tabela 1 – Dados observados no estudo de hepatite viral aguda.

Controle			Terapia		
Id	Tempo	Censura	Id	Tempo	Censura
1	1	0	16	1	1
2	2	0	17	1	1
3	3	1	18	1	1
4	3	1	19	1	0
5	3	0	20	4	0
6	5	0	21	5	1
7	5	0	22	7	1
8	16	0	23	8	1
9	16	0	24	10	1
10	16	0	25	10	0
11	16	0	26	12	0
12	16	0	27	16	0
12	16	0	28	16	0
14	16	0	29	16	0
15	16	0			

A Tabela 2 apresenta os respectivos cálculos da curva de sobrevivência Kaplan-Meier para os grupos controle e terapia dos dados analisados, seguindo as etapas descritas na seção 3.2. A biblioteca SURVIVAL do software R [46] efetua estes cálculos, a Figura 2 apresenta o gráfico das curvas de sobrevivência correspondentes aos resultados apresentados na Tabela 2 e calculados no R.

Algumas interpretações das curvas de sobrevivência são apresentadas a seguir: a função de sobrevivência, isto é, a probabilidade de sobreviver a três semanas no grupo controle é 84,6% enquanto que a no grupo com terapia é 78,6%. O grupo controle apresenta probabilidade de sobrevivência maior que o grupo com terapia em todos os tempos, assim pode-se supor que a terapia testada não é benéfica.

A aplicação do modelo de Cox sobre os dados analisados inicia-se considerando que a covariável é o indicador do grupo, onde o grupo controle assume o valor 0 e o grupo com a terapia de esteroide assume o valor 1. Considera-se, adicionalmente, duas taxas de falha $\lambda_0(t)$ e $\lambda_1(t)$ para os grupos 0 e 1, respectivamente.

Tabela 2 – Cálculo das curvas de sobrevivência no estudo de hepatite viral aguda para grupos controle e terapia.

Controle				Terapia			
t_i	d_i	n_i	$\hat{S}(t_i)$	t_i	d_i	n_i	$\hat{S}(t_i)$
3	2	13	0,846	1	3	14	0,786
				5	1	9	0,698
				7	1	8	0,611
				8	1	7	0,524
				10	1	6	0,437

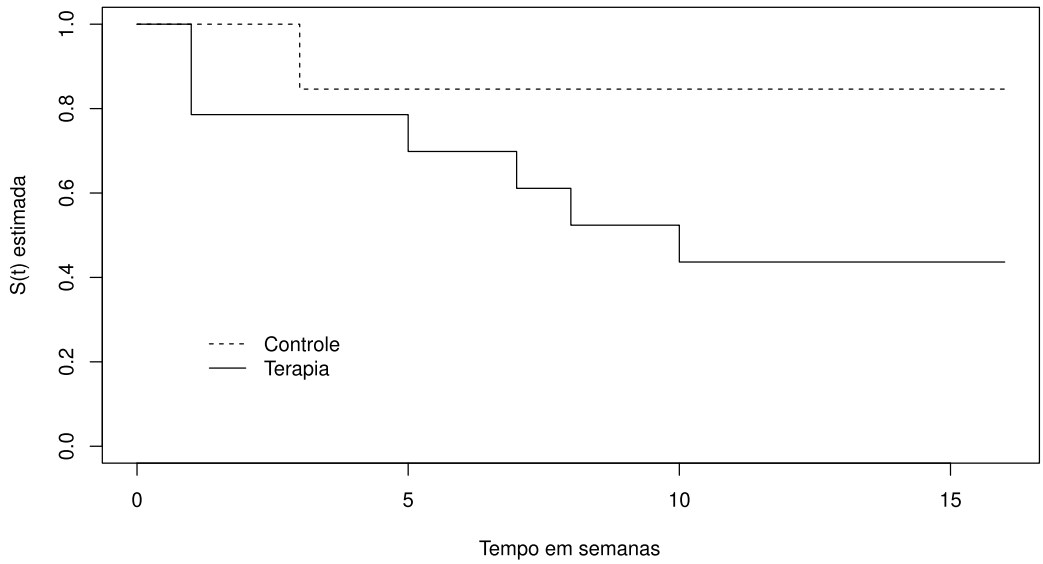

Figura 2 – Curvas de sobrevivência para grupos controle e terapia.

Assume-se uma proporcionalidade entre as duas taxas, isto é, $k = \lambda_1(t)/\lambda_0(t)$, onde k é uma razão das taxas de falha constante para todo t, também chamada de risco relativo.

Considerando:

$$x = \begin{cases} 0 \text{ se é grupo controle} \\ 1 \text{ se é grupo com terapia} \end{cases} \qquad e \qquad k = exp\{\beta x\},$$

temos:

$$\lambda(t) = \begin{cases} \lambda_1(t) = \lambda_0(t) exp\{\beta\} & \text{se} \quad x = 1 \\ \lambda_0(t) & \text{se} \quad x = 0 \end{cases}$$

A biblioteca SURVIVAL do software R [46] também é usada para a obtenção do modelo de Cox. Neste exemplo, o intuito é apresentar a formalização da construção do modelo de Cox e não os resultados, pois busca-se observar as diferenças com o modelo Buckley-James que será apresentado na próxima seção.

3.5 MODELO BUCKLEY-JAMES (BJ)

O modelo Buckley-James é definido como:

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{3.5}$$

onde $Y_i = ln(T_i)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})^\top$ é um vetor de parâmetros desconhecidos de dimensão p, $\mathbf{x}_i = (1, x_{1i}, ..., x_{1(p-1)})^\top$ é a i-ésima linha da matriz de desenho $\mathbf{X}$ de dimensão $n \times p$, $n > p$, e os erros aleatórios ϵ_1,...,ϵ_n são independentes e identicamente distribuídos como $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, para $i = 1, \ldots, n$, onde σ^2 é um parâmetro desconhecido maior que zero.

Considerando que a amostra tem tamanho n, a representação dos dados de sobrevivência da i-ésima observação, $i = 1, .., n$, é (t_i,δ_i,$\mathbf{x}_i$), onde $t_i = min(T_i, C_i)$ com T_i o tempo de sobrevivência e C_i o tempo de censura, $\delta_i = I(T_i \leq C_i)$ é o indicador da censura, $\mathbf{x}_i$ é o vetor (de dimensão p) de covariáveis medidas para a observação i.

Pode-se considerar que o modelo BJ é do tipo AFT (tempo de falha acelerado) tradicional, que estabelece a relação linear entre o logaritmo do tempo de sobrevivência e as covariáveis.

Quando não há censura, o modelo (3.5) é um modelo tradicional de regressão linear múltipla. Portanto, o coeficiente de regressão pode ser estimado pelo métodos de mínimos quadrados:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \qquad e \qquad \widehat{\sigma^2} = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.6}$$

No entanto, no contexto de análise de sobrevivência, o método dos mínimos quadrados e outros modelos de regressão comuns não podem ser implementados diretamente, devido à censura existente nos dados. Buckley e James propuseram estimar as observações censuradas por sua esperança condicional, dada a observação do logaritmo do tempo de censura correspondente T_i e das covariáveis $\mathbf{x}_i$:

$$y_i^* = \delta_i y_i + (1 - \delta_i) E(T_i | T_i > y_i, \mathbf{x}_i), i = 1, .., n \tag{3.7}$$

Numa observação sem censura $y_i^* = y_i$, isto é, o verdadeiro valor do logaritmo do tempo de sobrevivência. Quando a observação é censurada, o logaritmo do verdadeiro tempo de sobrevivência é obviamente maior que o observado e portanto, é adequado estimar T_i da observação censurada considerando sua esperança condicional, dado o tempo de censura correspondente e as covariáveis $E(T_i | T_i > y_i, \mathbf{x}_i)$. Teoricamente é possível provar que $E(Y_i^*) = E(T_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$ o que garante que y_i^* é um estimador não enviesado para T_i.

No modelo BJ, pelo exposto anteriormente, a esperança condicional $E(T_i|T_i > y_i, \mathbf{x}_i)$ é igual a $E(\mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i|\mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i > y_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + E(\epsilon_i|\epsilon_i > y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$. Portanto, esta pode ser calculada como:

$$E(T_i|T_i > y_i, \mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \int_{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}^{\infty} \frac{tdF(t)}{1 - F(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})}, \tag{3.8}$$

onde $F(.)$ é a função de distribuição acumulada do erro aleatório $\epsilon_i (\epsilon_i = y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$.

Supondo que o vetor de coeficientes de regressão $\boldsymbol{\beta}$ foi estimado como $\widehat{\boldsymbol{\beta}}$, então pode-se obter valores de resíduos observados com censura (ϵ_i, δ_i), onde $\epsilon_i = y_i - \mathbf{x}_i^\top \boldsymbol{\beta}$ para $i = 1, ..., n$. Com base nestes resíduos observados, pode-se estimar a função de distribuição $F(.)$ pelo método Kaplan-Meier (KM).

De acordo com o método KM, é preciso ordenar os resíduos observados de forma a ter $\epsilon_1 < \epsilon_2 < ... < \epsilon_n$, e reordenar os dados de sobrevivência observados $(y_i, \delta_i, \mathbf{x}_i)$, para $i = 1, ..., n$ de acordo com a ordem de classificação dos resíduos observados.

Considerando a função de sobrevivência denotada e definida como $S(.) = 1 - F(.)$, uma estimativa $\widehat{S(.)}$ pode ser obtida pelo método KM. Consequentemente, a esperança condicional do logaritmo do tempo de sobrevivência pode ser calculada por:

$$E(T_i|T_i > y_i, \mathbf{x}_i) = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} + \widehat{S}(\epsilon_i)^{-1} \sum_{\epsilon_j > \epsilon_i} \epsilon_j \delta_j \boldsymbol{\Delta} \widehat{S}(\epsilon_j), i = 1, ..., n, \tag{3.9}$$

onde $\boldsymbol{\Delta}\widehat{S}(\epsilon_j)$ é o valor do salto da função estimada $\widehat{S}(.)$ no resíduo ϵ_j. Portanto, y_i^* pode ser calculado por:

$$y_i^* = \delta_i y_i + (1 - \delta_i) \left[\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} + \widehat{S}(\epsilon_i)^{-1} \sum_{\epsilon_j > \epsilon_i} \epsilon_j \delta_j \boldsymbol{\Delta} \widehat{S}(\epsilon_j) \right], i = 1, .., n. \tag{3.10}$$

Depois de estimar todos os T_i censurados por y_i^*, o tempo de sobrevivência de todas as observações encontram-se sem censura. Assim, os coeficientes de regressão e a variância podem ser estimado segundo (3.6), considerando $Y^* = (y_1^*, y_2^*, ..., y_n^*)$.

No método BJ, o vetor de coeficientes de regressão precisa ser estimado iterativamente, pois os valores y_i^* são calculados sob a condição de que os $\boldsymbol{\beta}$ são conhecidos, enquanto os coeficientes de regressão $\boldsymbol{\beta}$ sob a condição de que os y_i^* são conhecidos.

3.6 A DISTRIBUIÇÃO t-STUDENT

De acordo com Casella et al. [60], considere uma amostra aleatória $X_1, X_2, \ldots, X_n$ proveniente de uma população normal com média μ e variância desconhecida σ^2. A estatística

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

segue uma distribuição *t-Student* com $n-1$ graus de liberdade, onde $\bar{X}$ é a média amostral e S é o desvio padrão amostral, definido a partir da variância amostral $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Esse resultado evidencia o papel fundamental da distribuição *t-Student* na inferência estatística, sendo especialmente relevante em situações de amostras pequenas e variância populacional desconhecida. A distribuição é amplamente utilizada na construção de intervalos de confiança e na realização de testes de hipótese. Além disso, a distribuição *t-Student* pode ser obtida como a razão entre uma variável normal padrão e a raiz quadrada de uma variável qui-quadrado independente dividida por seus graus de liberdade.

Algumas propriedades importantes desta distribuição são:

- **Simetria**: é simétrica em torno de zero.
- **Caudas pesadas**: possui caudas mais espessas que a distribuição normal, tornando-a mais conservadora em testes de hipóteses.
- **Convergência**: quando $\nu \to \infty$, converge para a normal padrão.
- **Média**: igual a zero para $\nu > 1$.
- **Variância**: igual a $\frac{\nu}{\nu-2}$ para $\nu > 2$; indefinida para $\nu \leq 2$.
- **Moda e mediana**: ambas localizadas em zero.

Observe que muitas dessas propriedades são compartilhadas com a distribuição normal padrão. As caudas mais pesadas estão associadas ao número de graus de liberdade, e a distribuição se aproxima da normal à medida que esse número aumenta. Isso faz com que a distribuição *t-Student* seja frequentemente utilizada em modelos robustos, como alternativa à normal em presença de *outliers*.

Para a construção de modelos robustos, um caminho usado é considerar às distribuições normal e *t-Student* como casos particulares dentro da classe das Distribuições de Mistura de Escala Normal (MEN), proposta por Andrews e Mallows (1974) [56] . Essa classe abrange distribuições simétricas que compartilham uma estrutura comum: são construídas como misturas de distribuições normais com diferentes escalas (variâncias), o que permite modelar caudas mais pesadas ou mais leves conforme necessário. Huaira Contreras (2014) [57] discute com detalhe esta classe de distribuições e apresenta uma aplicação para modelos com ponto de mudança.

Essa perspectiva unificada permite compreender a *t-Student* como uma extensão natural da normal, adaptada para cenários com maior variabilidade e menor informação sobre a população.

A estrutura MEN admite uma representação estocástica útil para simulação e interpretação:

$$Y = \mu + \kappa^{1/2}(U)Z, \tag{3.11}$$

onde:

- $Z \sim \mathcal{N}(0, \sigma^2)$ é uma variável normal padrão,
- U é uma variável aleatória positiva com distribuição $H(u; \boldsymbol{\nu})$, independente de Z,
- $\kappa(\cdot)$ é uma função de ponderação positiva.

Essa representação mostra que a variável Y é construída como uma normal com variância aleatória, controlada por U. No caso da distribuição *t-Student*, essa variância aleatória reflete a incerteza sobre a variância populacional, enquanto na distribuição normal ela é fixa. Assim, dentro da classe MEN, as distribuições normal e *t-Student* podem ser descritas como:

- **Distribuição Normal** $N(\mu, \sigma^2)$: Representa o caso mais simples da classe MEN, onde a função de ponderação é constante $\kappa(u) = 1$ e a variável U é degenerada, ou seja, assume valor fixo. Isso implica que não há variabilidade na escala, resultando em uma distribuição com caudas finas e comportamento padrão. A normal é, portanto, uma MEN com estrutura determinística. O momento condicional é $q(d) = 1$.
- **Distribuição t-Student** $t(\mu, \sigma^2, \nu)$: Representa a distribuição com ν graus de liberdade. Caracteriza-se por $\kappa(u) = 1/u$ e $U \sim \text{Gamma}(\nu/2, \nu/2)$, o que introduz variabilidade na escala. Essa estrutura gera caudas mais pesadas, tornando a *t-Student* mais robusta em presença de *outliers* e incertezas na variância populacional. A distribuição t surge como uma mistura de normais com variâncias aleatórias, controladas pela variável U. O momento condicional é $q(d) = (\nu - 1)/(\nu - d)$.

3.7 MODELO DE REGRESSÂO t-STUDENT

O modelo de regressão *t-Student* é definido como

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{3.12}$$

onde $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})^\top$ é um vetor de parâmetros desconhecidos de dimensão p, $\mathbf{x}_i = (1, x_{1i}, ..., x_{1(p-1)})^\top$ é a i-ésima linha da matriz de desenho $\mathbf{X}$ de dimensão $n \times p, n > p$ e, os erros aleatórios ϵ_1,...,ϵ_n são independentes e identicamente distribuídos como $\epsilon_i \overset{iid}{\sim} t(0, \sigma^2, \nu)$, para $i = 1, \ldots, n$, onde σ^2 é um parâmetro desconhecido maior que zero e ν são os graus de liberdade da distribuição *t-Student*, considerando a distribuição dentro da classe MEN.

Sugere-se que os graus de liberdade ν sejam considerados conhecidos e uma avaliação de vários possíveis valores de ν deverá ser feita para definir o modelo. Quando $\nu \to \infty$ a distribuição *t-Student* converge à normal.

A estimação dos parâmetros pelo estimador de máxima verossimilhança é dada por

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{P}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{P}\mathbf{Y}, \qquad \widehat{\sigma^2} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{P}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \tag{3.13}$$

onde

$$\mathbf{P} = diag(p_1, ..., p_n), \qquad p_i = \frac{\nu + 1}{\nu + d_i} \qquad d_i = \frac{(Y_i - \mathbf{x_i}^\top \widehat{\boldsymbol{\beta}^\top})^2}{\widehat{\sigma^2}} \quad i = 1, \ldots, n.$$

Observa-se que o cálculo de d_i requer os valores de $\boldsymbol{\beta}$ e σ^2 estimados, indicando que a solução do modelo requer do uso de um método iterativo. Finalmente, o modelo de regressão *t-Student* é uma alternativa ao modelo de regressão normal para modelar de forma mais robusta conjuntos com presença de dados extremos. Desta forma, é possível adicionar dentro do modelo BJ os conceitos do modelo *t-Student* e oferecer uma alternativa que possa lidar melhor com dados extremos.

3.8 TESTE RESET DE RAMSEY

O *Regression Equation Specification Error Test* (RESET), proposto por Ramsey (1969) [58], é um teste estatístico utilizado para avaliar a adequação da especificação funcional de um modelo de regressão linear. Foi proposto para detectar se o modelo linear estimado omite variáveis relevantes ou se a forma funcional escolhida não captura corretamente a relação entre as variáveis independentes e a variável dependente.

Assim, o teste RESET avalia as seguintes hipóteses:

- H_0: O modelo está corretamente especificado. Os termos adicionais $(\hat{y}^2, \hat{y}^3, \dots)$ não são significativos.
- H_1: O modelo está mal especificado. Os termos adicionais são significativos, indicando que a forma funcional original não captura toda a relação entre as variáveis.

Para testar as hipóteses propostas segue-se o seguinte raciocínio, se o modelo linear está corretamente especificado, então os valores ajustados $\hat{y}$ não devem conter informação adicional que explique a variável dependente além das covariáveis originais. Para verificar isso, o teste adiciona ao modelo original termos polinomiais dos valores ajustados, como $\hat{y}^2$, $\hat{y}^3$, etc., e testa se esses termos são estatisticamente significativos.

Desta forma, ao rejeitar H_0 sugere-se que o modelo linear pode estar omitindo variáveis ou necessitar de transformações não lineares, consequentemente sugere-se a adoção de um modelo não linear.

O teste RESET encontra-se implementado no software R [46] como a função `resettest()` do pacote `lmtest`. Os passos seguidos para realizar o teste são:

1. Estima-se o modelo linear original com `lm()`.
2. Calculam-se os valores ajustados $\hat{y}$.
3. Adicionam-se ao modelo os termos polinomiais de $\hat{y}$.
4. Testa-se se os coeficientes desses termos são diferentes de zero.

Zeiles et al. (2002) [59] descreve a implementação do teste RESET no pacote `lmtest`) do software R. Por configuração padrão, o teste adiciona termos quadráticos e cúbicos ao modelo, isto é, $\hat{y}^2$ e $\hat{y}^3$, de modo a verificar a presença de efeitos desses graus. Logo, o teste avalia simultaneamente a existência de efeitos quadráticos e cúbicos dos valores ajustados. É possível avaliar outros graus polinomiais, basta utilizar o argumento `power`, definindo os graus que se deseja avaliar. Para rejeitar H_0, isto é, existir evidência estatística suficiente para considerar que um modelo não linear é o adequado, basta que um efeito seja significativo.

Como um exemplo apresentamos um código simples que descreve o uso do teste RESET no sofware R para validar polinômios até o grau 4, para isto consideramos um conjunto dados denominado Dados.Teste que tem as variáveis y, $x1$ e $x2$. Assim temos

```
library(lmtest)

# Modelo linear
modelo <- lm(y ~ x1 + x2, data = dados.Teste)

# Teste RESET com configuração até grau 4 (power = 2:4)
resettest(modelo, power=2:4)
# Teste RESET com configuração default (power = 2:3)
```

A seguir apresenta-se a formalização do teste.

Considerando o modelo linear:

$$y = X\beta + \varepsilon$$

onde X é a matriz de covariáveis e ε é o erro. O teste RESET estima o modelo ampliado:

$$y = X\beta + \gamma_1\hat{y}^2 + \gamma_2\hat{y}^3 + \cdots + \varepsilon$$

O teste RESET é baseado na estatística F de Snedecor. Este compara o modelo restrito (sem termos adicionais) com o modelo ampliado. A estatística de teste é dada por:

$$F = \frac{(SQR_r - SQR_a)/q}{RSS_a/(n-k-q)}$$

onde SQR_r é a soma dos quadrados dos resíduos do modelo restrito, SQR_a é a soma dos quadrados dos resíduos do modelo ampliado, q é o número de termos adicionais e $n-k-q$ são os graus de liberdade. Se F for significativo, rejeita-se H_0.

3.9 MÁQUINA DE APRENDIZADO EXTREMO (ELM)

A Máquina de Aprendizado Extremo (ELM) (Huang et al, 2004 e 2006 [12] [13]) é uma rede neural do tipo Rede Neural Feedforward de Camada Única (SLFN). Se diferencia dos algoritmos tradicionais de aprendizagem baseados em gradiente para SLFN, pois os parâmetros da camada oculta (pesos e termos de viés) são atribuídos aleatoriamente, sem ajuste, enquanto os pesos da camada de saída são determinados encontrando a solução por mínimos quadrados. Por essa razão, é um algoritmo de aprendizado rápido, apresentando vantagens como bom desempenho de generalização, maior rapidez de treinamento em comparação às redes treinadas por backpropagation e desempenho competitivo frente às Máquinas de Vetor de Suporte (SVM).

O uso do ELM é diversificado. No aprendizado de máquina supervisionado, o ELM é aplicado para a resolução de problemas de classificação e regressão. Já no aprendizado não supervisionado, o ELM pode ser utilizado para resolver problemas de clusterização e redução de dimensionalidade.

Para um problema de regressão, considera-se uma amostra de treinamento de tamanho n, onde a i-ésima observação é dada por $(\mathbf{x}_i, y_i)$,sendo $\mathbf{x}_i$ de dimensão p e y_i de dimensão 1 ($i = 1, 2, ..., n$). Uma rede neural SLFN com uma função de ativação $g(.)$ e Q neurônios ocultos é definida como:

$$f_Q(\mathbf{x}_i) = \sum_{q=1}^{Q} g(\mathbf{x}_i^\top \mathbf{w}_q + b_q)\alpha_q = \mathbf{G}(\mathbf{x}_i^\top \mathbf{W} + \mathbf{B})\boldsymbol{\alpha}, i = 1, ..., n \tag{3.14}$$

onde $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_Q)$ e $\mathbf{B} = (b_1, b_2, ..., b_Q)$ representam os pesos e termos de viés da camada oculta, $\mathbf{G}(\mathbf{x}_i^\top \mathbf{W} + \mathbf{B}) = (g(\mathbf{x}_i^\top \mathbf{w}_1 + b_1), g(\mathbf{x}_i^\top \mathbf{w}_2 + b_2), ..., g(\mathbf{x}_i^\top \mathbf{w}_Q + b_Q))$ é o vetor de saída de dimensão Q da camada oculta em relação a $\mathbf{x}_i$, e $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_Q)^\top$ é o vetor de pesos de saída que conecta a camada oculta à camada de saída. Os parâmetros de camada oculta $\mathbf{W}$ e $\mathbf{B}$ são gerados aleatoriamente a partir de duas funções arbitrárias de distribuição de probabilidade contínua considerando dimensões p e 1, respectivamente.

Além disso, considerando a matriz de entrada $\mathbf{X}$ e a matriz de destino $\mathbf{Y}$ da rede:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ \vdots & \vdots & ... & \vdots \\ x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix} \quad e \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \tag{3.15}$$

A matriz de saída da camada oculta é:

$$\mathbf{H} = \begin{bmatrix} G(\mathbf{x}_1^\top \mathbf{W} + \mathbf{B}) \\ G(\mathbf{x}_2^\top \mathbf{W} + \mathbf{B}) \\ \vdots \\ G(\mathbf{x}_n^\top \mathbf{W} + \mathbf{B}) \end{bmatrix} = \begin{bmatrix} g(\mathbf{x}_1^\top \mathbf{w}_1 + b_1) & g(\mathbf{x}_1^\top \mathbf{w}_2 + b_2) & ... & g(\mathbf{x}_1^\top \mathbf{w}_Q + b_Q) \\ g(\mathbf{x}_2^\top \mathbf{w}_1 + b_1) & g(\mathbf{x}_2^\top \mathbf{w}_2 + b_2) & ... & g(\mathbf{x}_2^\top \mathbf{w}_Q + b_Q) \\ \vdots & \vdots & ... & \vdots \\ g(\mathbf{x}_n^\top \mathbf{w}_1 + b_1) & g(\mathbf{x}_n^\top \mathbf{w}_2 + b_2) & ... & g(\mathbf{x}_n^\top \mathbf{w}_Q + b_Q) \end{bmatrix} \tag{3.16}$$

Assim, a partir de (3.14) e (3.16), a rede neural SLFN pode ser formulada como

$$f_Q(\mathbf{X}) = \mathbf{H}\boldsymbol{\alpha}. \tag{3.17}$$

Como os parâmetros da camada oculta $\mathbf{W}$ e $\mathbf{B}$ são gerados aleatoriamente, a matriz de saída $\mathbf{H}$ pode ser determinada. Em seguida, o vetor de pesos de saída $\boldsymbol{\alpha}$ pode ser estimado resolvendo uma solução de mínimos quadrados, utilizando a pseudo-inversa de Moore-Penrose:

$$\hat{\boldsymbol{\alpha}} = \mathbf{H}^\dagger \mathbf{Y} = \begin{cases} (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{Y}, & n < Q \\ \mathbf{H}^\top (\mathbf{H}\mathbf{H}^\top)^{-1} \mathbf{Y}, & n \geq Q \end{cases}. \tag{3.18}$$

Em (3.18), $\mathbf{H}^\dagger$ denota a inversa generalizada de Moore–Penrose da matriz $\mathbf{H}$. Quando o número de neurônios ocultos é maior que o número de amostras de treinamento, $n < Q$, a matriz H de ordem $n \times Q$ é uma matriz de posto completo em colunas. Consequentemente, a matriz $\mathbf{H}^\top \mathbf{H}$ de ordem $Q \times Q$ é invertível. De acordo com a definição da inversa generalizada de Moore–Penrose, a matriz $(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}$ é a inversa generalizada de Moore–Penrose da matriz $\mathbf{H}$. De forma similar, quando $n \geq Q$, a matriz $\mathbf{H}$ é uma matriz de posto completo em linhas e a matriz $\mathbf{H}\mathbf{H}^\top$ de ordem $n \times n$ é invertível. A matriz $\mathbf{H}^\top (\mathbf{H}\mathbf{H}^\top)^{-1}$ é a inversa generalizada de Moore–Penrose da matriz $\mathbf{H}$.

3.10 ALGORITMO BOOSTING L2

Um algoritmo de boosting que ajusta iterativamente o vetor do gradiente negativo por meio de um procedimento base é essencialmente um algoritmo de *Functional Gradient Descent* (FGD). Friedman (2001) [49] apresentou uma estrutura geral de aprendizado para o algoritmo FGD.

Na estrutura geral de aprendizado do algoritmo FGD, a seleção de diferentes funções de perda $l(y, f(x))$ para boosting pode gerar vários algoritmos de boosting correspondentes.

O algoritmo de Boosting L2 é o mais simples para regressão, utilizando o erro quadrático $(y - f(x))^2$ como função de perda. Ao usar a função de perda do erro quadrático, o vetor do gradiente negativo é igual ao vetor residual ordinário. Portanto, no algoritmo de Boosting L2, ajustamos o vetor residual ordinário por meio de um modelo base a cada iteração. Ao aplicar o algoritmo de Boosting L2 para resolver problemas de regressão, precisamos escolher um regressor base (por exemplo, regressão linear simples).

Em termos simples, um *modelo fraco* é um modelo de previsão muito básico, que sozinho não consegue explicar toda a complexidade dos dados. Ele pode ser, por exemplo, uma regressão linear simples ou uma árvore de decisão muito rasa. Embora cada modelo fraco tenha desempenho limitado, o algoritmo de boosting combina muitos desses modelos em sequência, de forma que cada um corrige os erros do anterior. O resultado final é um *modelo forte*, capaz de alcançar alta precisão ao aproveitar a força coletiva de vários modelos fracos.

Assim, o conceito fundamental por trás do Boosting L2 é a minimização da função de perda quadrática, onde o modelo é ajustado iterativamente. A cada iteração, um novo modelo é treinado para prever os resíduos do modelo anterior. Isso é feito utilizando uma combinação linear de modelos fracos. O Boosting L2 ajusta os coeficientes dos modelos fracos de forma a minimizar a função de perda global, empregando técnicas de otimização como o gradiente descendente. O **Algoritmo 1** ilustra a abordagem descrita.

Algorithm 1: Algoritmo Boosting L2

Data: Dados de treino (X, y),
número de aprendizes base M,
taxa de aprendizado γ

Result: Modelo final $f_M(x)$

Inicializar o modelo $f_0(x) = \overline{y}$;
for $m = 1$ ***até*** M **do**
    Calcular os resíduos $e_i^{(m)} = y_i - f_{m-1}(x_i)$;
    Ajustar o modelo base $u_m(x)$ nos resíduos $e_i^{(m)}$;
    Atualizar o modelo: $f_m(x) = f_{m-1}(x) + \gamma u_m(x)$;
Retornar o modelo final $f_M(x)$;

O estudo de Friedman [49] mostra que, no algoritmo de Boosting L2, a seleção da taxa de aprendizado γ tem pouco impacto, desde que seja escolhida pequena o suficiente, como $\gamma = 0.1$. O número de regressores base M é um parâmetro de ajuste no algoritmo, que pode ser determinado, por exemplo, por um esquema de validação cruzada (CV).

3.11 DESENVOLVIMENTO METODOLÓGICO

Nesta seção são apresentados os conceitos e as articulações que fundamentam a metodologia proposta. O desenvolvimento metodológico aqui exposto faz uso de ferramen-

tas estatísticas e de aprendizado de máquina. Tais conceitos, além de consistentes, são amplamente conhecidos e consolidados, garantindo que a proposta esteja fundamentada em bases teóricas e práticas reconhecidas.

3.11.1 Formulação do modelo proposto

Considerando que se propõe um modelo adaptativo capaz de avaliar a presença ou ausência de linearidade, incorporando robustez frente a observações extremas em dados de sobrevivência, e que tal abordagem se fundamenta no modelo de Buckley-James aliado a um comitê L2 Boosting de ELM, descrevem-se, a seguir, as articulações que sustentam a proposta.

3.11.1.1 Determinação da estrutura linear/não linear

O teste RESET foi introduzido por James B. Ramsey em 1969, em um contexto em que a econometria buscava métodos mais robustos para avaliar a validade dos modelos lineares clássicos. A preocupação central residia no fato de que tais modelos frequentemente omitiam variáveis relevantes ou não capturavam adequadamente relações não lineares, o que podia conduzir a inferências incorretas. Desde então, o RESET consolidou-se como uma ferramenta padrão nos diagnósticos de regressão, sendo amplamente empregado em aplicações econométricas. Situação análoga ocorre em dados da área médica, em especial em estudos de sobrevivência, nos quais os modelos mais utilizados também são lineares.

Em situações práticas, em que a forma funcional da relação entre variáveis não é evidente, como em estudos longitudinais, pode-se observar a necessidade de incorporar efeitos não lineares do tempo ou de variáveis biomédicas. Especificamente, em modelos de sobrevivência, embora o RESET tenha sido originalmente desenvolvido para regressão linear, sua utilização também se justifica em modelos de Buckley-James, uma vez que, por construção, estes são formulados como modelos de regressão com dados censurados.

Cabe destacar que o teste RESET não indica qual variável é não linear. Este apenas sinaliza que o modelo linear não é consistente. Para esta situação, o uso do ELM avalia conjuntamente todas as covariáveis e consequentemente, e uma vez que o teste RESET indique não linearidade, o ELM mostra capacidade de tratar relações não lineares de forma eficiente.

É importante salientar que o teste RESET não permite identificar quais covariáveis apresentam especificações não lineares; sua função é apenas indicar que a forma linear do modelo é insuficiente. Diante dessa limitação, a utilização do ELM torna-se adequada, pois avalia simultaneamente todas as covariáveis e, uma vez que o RESET sinalize a presença de não linearidade, o ELM apresenta capacidade de modelar tais relações de maneira eficiente e abrangente.

Para a utilização de teste RESET na proposta avalia-se o valor p, considerando o seguinte critério para decisão:

- Se o valor-p do teste for menor que 0,05, rejeita-se H_0, indicando que o modelo pode estar mal especificado.
- Se o valor-p for maior ou igual a 0,05, não há evidência de erro de especificação.

Finalmente, algumas limitações do teste RESET são discutidas:

- **Baixo poder discriminativo em alguns cenários**: pode não detectar especificações incorretas se os termos polinomiais não capturam bem a forma funcional verdadeira. Nesta situação, o uso de ELM pode tratar outros tipos de relações não lineares. Por outro lado, relações quadráticas ou cúbicas explicam um grande número de relações não lineares.
- **Sensibilidade ao número de termos**: incluir muitos termos pode levar a sobreajuste ou resultados espúrios. Isto pode ser contornado fazendo uma análise preliminar para selecionar covariáveis significativas para o modelo ou incluindo métodos de regularização nos modelos.

O teste RESET constitui uma ferramenta robusta para avaliar a adequação da especificação de modelos lineares. Sua implementação em R é direta e, por padrão, incorpora termos quadráticos e cúbicos dos valores ajustados. Em aplicações práticas, como em estudos longitudinais ou em modelos de regressão de maior complexidade, o RESET pode ser utilizado como parte do processo de validação, conferindo maior consistência à interpretação dos resultados. Ademais, a integração com métodos baseados em aprendizado de máquina, como o ELM, potencializa sua aplicação ao oferecer maior capacidade de captura de relações não lineares.

3.11.1.2 Inclusão de robustez no modelo

A distribuição *t-Student* constitui uma alternativa robusta à normal, pois incorpora caudas mais pesadas e reduz a influência de observações extremas (*outliers*) nos procedimentos de inferência. Essas características a tornam particularmente adequada para modelagem robusta em cenários com dados extremos.

Na classe MEN, as distribuições normal e *t-Student* podem representar diferentes comportamentos de cauda, mantendo a simetria e a estrutura normal como base. Além disso, por possuírem formas fechadas e distribuições conhecidas para U, são especialmente adequadas para aplicações computacionais e inferência estatística.

A Figura 3 mostra que, para distribuições normais com diversos valores de média e variância, é possível encontrar alternativas de distribuições *t-Student* com caudas pesadas,

considerando as definições da classe MEN. Observa-se que a simetria é mantida, mas as formas — especialmente nas caudas — apresentam diferenças que dependem dos parâmetros próprios destas distribuições. A convergência para a distribuição normal ocorre no limite em que $\nu \to \infty$, reforçando a ideia de que a *t-Student* constitui uma generalização robusta da normal.

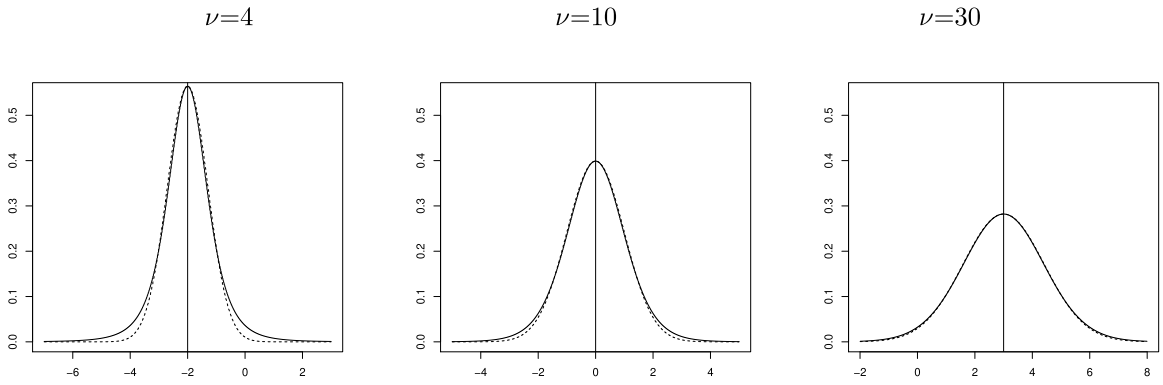

Figura 3 – Algumas distribuições t-Student (linha cheia) como alternativas para $N(-2, 0.5)$, $N(0, 1)$ e $N(3, 2)$ (linha pontilhada).

Dessa forma, a incorporação da distribuição *t-Student* fortalece a especificação dos modelos, ao oferecer flexibilidade para acomodar diferentes níveis de variabilidade e assegurar maior consistência estatística em aplicações práticas.

A robustez conferida pela distribuição *t-Student* pode ser incorporada de forma natural em modelos de sobrevivência, em particular no modelo de Buckley–James, que consiste em uma extensão da regressão linear para dados censurados.

A integração com erros *t-Student* amplia a capacidade do modelo Buckley–James ao lidar com observações extremas. Como discutido anteriormente, a distribuição *t-Student* possui caudas mais pesadas que a normal, o que reduz a influência de *outliers* e torna os estimadores mais robustos. Dessa forma, ao substituir a suposição de erros normais por erros *t-Student*, o modelo Buckley–James passa a oferecer maior consistência estatística em cenários práticos, especialmente em estudos longitudinais ou biomédicos, nos quais a presença de dados aberrantes é comum.

Do ponto de vista computacional, essa integração é viável porque tanto o Buckley–James quanto a regressão com erros *t-Student* podem ser formulados em termos matriciais e resolvidos por métodos iterativos, como algoritmos de mínimos quadrados ponderados (WLS) ou esquemas EM/IRLS. A estimação dos parâmetros segue a mesma lógica: os resíduos censurados são imputados e, em seguida, ponderados de acordo com os pesos derivados da distribuição *t-Student*, resultando em estimativas robustas para os coeficientes e para a variância.

Em síntese, a combinação entre o modelo Buckley–James e a distribuição *t-Student*

constitui uma extensão robusta da análise de sobrevivência. Essa abordagem preserva a estrutura linear do modelo original, mas adiciona flexibilidade para lidar com caudas pesadas e observações extremas, tornando-se uma alternativa poderosa em aplicações médicas, econométricas e em outras áreas nas quais a censura e a não normalidade dos dados são características recorrentes.

3.11.1.3 Integração de ELM e Buckley–James com erros *t-Student*

O Buckley–James é uma extensão da regressão linear para dados censurados, mas sua formulação linear pode ser limitada quando a relação entre covariáveis e tempo de sobrevivência apresenta componentes não lineares. Nesse contexto, o ELM oferece flexibilidade ao permitir diferentes funções de ativação, como identidade para capturar linearidade ou sigmoide para modelar relações não lineares, ampliando a capacidade de representação do modelo.

Um dos principais desafios em modelos de sobrevivência é a presença de observações extremas (*outliers*), que podem distorcer estimativas e comprometer a inferência, o tratamento de dados extremos pode ser incorporado ao Buckley–James por meio da substituição da suposição de erros normais por erros *t-Student*. A distribuição *t-Student*, por possuir caudas mais pesadas, reduz a influência de observações aberrantes e confere maior robustez às estimativas. Dessa forma, o modelo Buckley–James com erros *t-Student* passa a oferecer maior consistência estatística em cenários práticos, especialmente em estudos biomédicos e longitudinais.

Algumas ideias propostas por Chen et al. (2017) [61] foram aproveitadas para o modelo Buckley–James robusto. O uso de pesos iterativos, como no método IRLS, permite que os resíduos sejam ponderados de forma adaptativa, atribuindo menor peso às observações extremas e garantindo que os coeficientes estimados não sejam dominados por *outliers*. Além disso, a integração com o ELM fornece flexibilidade para modelar linearidade ou não linearidade, enquanto o esquema de pesos derivados da distribuição *t-Student* assegura robustez contra caudas pesadas.

Em síntese, a integração proposta ocorre substituindo a matriz de covariáveis $\mathbf{X}$ do modelo Buckley–James pela matriz $\mathbf{H}$ obtida a partir da aplicação da Máquina de Aprendizado Extremo (ELM). Essa matriz $\mathbf{H}$ resulta da aplicação de funções de ativação sobre combinações lineares das covariáveis originais, permitindo ampliar o espaço de representação. Quando se utiliza a função de ativação identidade, o modelo preserva a linearidade; já com a função sigmoide, torna-se capaz de capturar relações não lineares entre as covariáveis e o tempo de sobrevivência.

A estimação dos parâmetros de saída $\boldsymbol{\alpha}$ é realizada por meio de um esquema iterativo de mínimos quadrados ponderados, em que os resíduos são reponderados de

acordo com a distribuição *t-Student*. A solução na iteração $k+1$ é dada por

$$\boldsymbol{\alpha}^{(k+1)} = \left(\mathbf{H}^\top \mathbf{P}^{(k)} \mathbf{H}\right)^{-1} \mathbf{H}^\top \mathbf{P}^{(k)} \mathbf{Y}, \tag{3.19}$$

onde $\mathbf{P}^{(k)} = \text{diag}(p_1, \ldots, p_n)$ é a matriz diagonal de pesos, com

$$p_i = \frac{\nu + 1}{\nu + d_i}, \qquad d_i = \frac{\left(Y_i - \mathbf{x_i}^\top \widehat{\boldsymbol{\beta}}\right)^2}{\widehat{\sigma^2}}, \quad i = 1, \ldots, n. \tag{3.20}$$

Dessa forma, o modelo Buckley–James com ELM e erros *t-Student* preserva a estrutura de análise de sobrevivência, mas adiciona flexibilidade para lidar com linearidade ou não linearidade e robustez contra observações extremas. Essa integração constitui uma alternativa poderosa em cenários práticos, como aplicações biomédicas, econométricas e longitudinais, nos quais censura, não normalidade e dados extremos coexistem.

3.12 Integração Boosting L2 com BJ-ELM- *t-Student*

A integração entre o modelo Buckley–James, a Máquina de Aprendizado Extremo (ELM) e a distribuição *t-Student* pode ser ampliada por meio do algoritmo de Boosting L2. Nesse contexto, o Boosting atua como um mecanismo de ajuste iterativo, em que cada iteração corresponde ao treinamento de um novo ELM. Assim, o modelo final é construído como uma combinação linear de múltiplos ELMs, construídos através da correção dos resíduos de forma sequencial.

O papel do ELM permanece central: a matriz de covariáveis $\mathbf{X}$ é substituída pela matriz $\mathbf{H}$, obtida a partir das funções de ativação aplicadas às combinações lineares das covariáveis originais. A função identidade preserva a linearidade, enquanto a função sigmoide permite capturar relações não lineares. Dessa forma, o Boosting L2 adiciona flexibilidade ao processo, permitindo que o modelo se adapte tanto a estruturas lineares quanto não lineares.

A robustez é garantida pela incorporação da distribuição *t-Student*, que confere menor peso às observações extremas por meio da matriz de ponderação $\mathbf{P}$. A cada iteração, os resíduos são recalculados e ponderados, reduzindo a influência de *outliers* e assegurando maior consistência estatística. O esquema iterativo do Boosting L2, aliado ao reponderamento via $\mathbf{P}$, resulta em estimativas mais estáveis e confiáveis.

Um resultado clássico mostra que o Boosting L2, quando aplicado com regressão linear como modelo de aprendizado base, não consegue evoluir além da solução única de mínimos quadrados ordinários. Seja $\mathbf{X}$ a matriz de covariáveis e $\mathbf{y}$ o vetor resposta. A solução de mínimos quadrados é dada por:

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

As previsões resultantes são:

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}, \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$

O vetor de resíduos é:

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y},$$

com a propriedade de ortogonalidade $\mathbf{X}^\top\mathbf{e} = \mathbf{0}$.

Se o segundo modelo de aprendizado base for também uma regressão linear, suas previsões seriam:

$$\begin{aligned}
\widehat{\mathbf{y}}_2 &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{e} \\
&= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{I} - \mathbf{H})\mathbf{y} \\
&= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y} \\
&= (\mathbf{H} - \mathbf{H})\mathbf{y} = \mathbf{0}.
\end{aligned}$$

Conclui-se que as previsões do segundo modelo são nulas. A regressão linear subsequente é incapaz de detectar qualquer padrão nos resíduos, pois estes, por construção, não possuem componente linear explicável pelas variáveis preditoras $\mathbf{X}$. Consequentemente, o processo iterativo de Boosting L2 torna-se ineficaz.

Essa limitação não se verifica quando utilizamos o *Extreme Learning Machine* (ELM) como modelo base. Diferentemente da regressão linear clássica, o ELM gera uma matriz de saída $\mathbf{H}$ construída a partir de funções de ativação e parâmetros $(\mathbf{W}, \mathbf{B})$ que podem ser redefinidos a cada iteração. Isso significa que, mesmo com a função identidade, novas matrizes $\mathbf{H}$ são geradas em cada passo, projetando os resíduos em subespaços distintos e evitando a ortogonalidade rígida que bloqueia o aprendizado sequencial.

Quando se utiliza a função sigmoide, o ganho é ainda maior, pois o espaço de representação é expandido de forma não linear, permitindo que cada nova matriz $\mathbf{H}$ capture padrões residuais complexos. Assim, tanto com identidade quanto com sigmoide, desde que novas matrizes $\mathbf{H}$ sejam geradas a cada iteração, o Boosting L2 com ELM não sofre da incompatibilidade fundamental observada na regressão linear. O resultado é um processo iterativo eficaz, capaz de lidar com censura, não linearidade e dados extremos de forma robusta.

3.13 Algoritmo Adaptativo BJ-ELM-*t-Student* com Boosting L2

O modelo BJ-ELM-*t-Student* pode ser estendido para uma versão adaptativa, na qual decisões fundamentais são tomadas automaticamente ao final de cada ciclo k do algoritmo. Essa adaptação confere maior flexibilidade ao processo, permitindo que o modelo se ajuste dinamicamente às características dos dados e otimize seu desempenho sem necessidade de intervenção manual.

Escolha Automática da Função de Ativação

A primeira escolha adaptativa refere-se à determinação da linearidade ou não do modelo. Para isso, aplica-se o teste RESET, conforme descrito na Subseção 3.11.1.1. O resultado do teste orienta a seleção da função de ativação utilizada na construção da matriz $\mathbf{H}$ do ELM. Se o teste indicar adequação de uma estrutura linear, utiliza-se a função identidade; caso contrário, adota-se uma função não linear, como a sigmoide. Importante destacar que essa escolha é realizada apenas ao final de cada ciclo k, quando os valores imputados $y^{*(k)}$ são atualizados segundo o conceito do modelo Buckley–James. Dessa forma, evita-se a incompatibilidade fundamental observada em regressões lineares simples, uma vez que novas matrizes $\mathbf{H}$ são geradas a cada ciclo, seja no espaço linear ampliado ou em espaços não lineares mais expressivos.

Atualização Adaptativa dos Graus de Liberdade

A segunda escolha adaptativa envolve a atualização do grau de liberdade ν da distribuição *t-Student*. Essa atualização é realizada ao final de cada ciclo k, por meio de um processo de otimização que consiste em maximizar a expressão:

$$\sum_i \log \mathbb{K}_i, \quad \text{onde} \quad \mathbb{K}_i = \frac{2^{\frac{1}{2}} \nu^{\frac{\nu}{2}} \Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) (d_i + \nu)^{\frac{\nu+1}{2}}},$$

sendo $\Gamma(x)$ a função gama avaliada em x. Esse procedimento busca o valor ótimo de ν, ajustando a robustez do modelo de acordo com a presença de observações extremas. Como os valores $y^{*(k)}$ são atualizados apenas ao final de cada ciclo, é nesse momento que a adaptação da robustez se torna consistente.

Flexibilidade e Automatização no Ciclo BJ

Essas escolhas automáticas — da função de ativação e do grau de liberdade — são realizadas ao final de cada ciclo k do algoritmo BJ-ELM com Boosting L2. O resultado é um processo adaptativo que combina linearidade ou não linearidade conforme necessário, ajusta a robustez estatística em função da estrutura dos dados e, ao mesmo tempo, aproveita o mecanismo iterativo do boosting para corrigir resíduos sucessivos dentro de cada ciclo. Em síntese, o BJ-ELM-*t-Student* adaptativo com Boosting L2 representa uma metodologia robusta e versátil, capaz de se ajustar dinamicamente às condições empíricas e oferecer ganhos adicionais de desempenho.

3.13.1 Algoritmo proposto

Como descrito em (3.5), o modelo BJ ajusta o logaritmo do tempo de sobrevivência, considerando uma relação linear com as covariáveis e assumindo uma distribuição normal dos erros do modelo. Entretanto, as aplicações reais de análise de sobrevivência

demonstram que os efeitos das covariáveis são frequentemente mais complexos. Relações de interação entre covariáveis e a presença de não linearidade são comuns, e a ocorrência de dados extremos pode comprometer a suposição de normalidade dos erros, resultando em degradação da capacidade preditiva do modelo BJ.

Considerando dados de sobrevivência com censura à direita, a proposta consiste em um algoritmo do tipo Boosting L2 baseado em ELM, com o objetivo de superar as dificuldades do modelo BJ em processar dados que apresentam interações ou não linearidade. Além disso, propõe-se a adoção da distribuição *t-Student* como alternativa à suposição de normalidade dos erros, conferindo maior robustez frente a observações extremas.

Primeiramente, o modelo proposto substitui a combinação linear de covariáveis do modelo BJ apresentado em (3.5) pela função de saída não linear obtida a partir da aplicação de um algoritmo Boosting L2 baseado em ELM, isto é:

$$Y_i = f(\mathbf{x}_i) + \epsilon_i, \tag{3.21}$$

onde $Y_i = \ln(T_i)$ e $f(\mathbf{x}_i)$ é a função de saída do modelo Boosting L2 baseada em ELM com M aprendizes base para o indivíduo i, $i = 1, 2, ..., n$. Assim, esta função é expressa como:

$$f(\mathbf{x}_i) = \sum_{m=1}^{M} G(\mathbf{x}_i^\top \mathbf{W}^{(m)} + \mathbf{B}^{(m)})\boldsymbol{\alpha}^{(m)}. \tag{3.22}$$

A cada iteração do método BJ, ajusta-se o logaritmo do tempo de sobrevivência por um modelo de reforço baseado em ELM com o mesmo número de covariáveis base em vez de um modelo de regressão linear múltipla, considerando o método de estimação proposto para um modelo *t-Student* que inclui a matriz de ponderação P como definido em (3.13). Essa matriz P é construída a partir das densidades da distribuição *t-Student*, atribuindo menor peso às observações com resíduos extremos. A atualização de ν é realizada por meio da maximização da soma dos logaritmos das densidades, garantindo que o modelo se ajuste dinamicamente ao nível de robustez necessário.

Logo, os valores y_i^*, para $i = 1, 2, ..., n$, são calculados por:

$$y_i^* = \delta_i y_i + (1 - \delta_i)\left[\widehat{f}(\mathbf{x}_i) + \widehat{S}(\epsilon_i)^{-1} \sum_{\epsilon_j > \epsilon_i} \epsilon_j \delta_j \boldsymbol{\Delta} \widehat{S}(\epsilon_j)\right], \tag{3.23}$$

onde $\epsilon_1 < \epsilon_2 < ... < \epsilon_n$ são resíduos observados ordenados e calculados por $\epsilon_i = y_i - \widehat{f}(\mathbf{x}_i)$.

Para incluir flexibilidade no algoritmo proposto, de forma que relações de interação entre covariáveis e a presença de não linearidade sejam melhor tratadas, duas funções de ativação $g(.)$ são utilizadas: a função identidade e a função sigmoide, definidas como:

$$g(x) = x, \qquad g(x) = \frac{1}{1 + e^{-x}}. \tag{3.24}$$

O processo adaptativo ocorre ao final de cada ciclo k do algoritmo BJ: aplica-se o teste RESET sobre os valores $y^{*(k)}$ para decidir se a função de ativação permanece identidade ou se deve ser atualizada para sigmoide, e simultaneamente otimiza-se o grau de liberdade ν da distribuição *t-Student*. Dessa forma, o algoritmo combina o mecanismo iterativo do Boosting L2 com decisões automáticas de linearidade e robustez, evitando a incompatibilidade fundamental da regressão linear e ampliando a capacidade de generalização frente a dados complexos e extremos.

Em síntese, o modelo BJ-ELM-*t-Student* adaptativo combina o mecanismo iterativo do Boosting L2 com decisões automáticas de linearidade/não linearidade e robustez, superando as limitações do modelo BJ clássico e ampliando sua capacidade de generalização. A implementação foi realizada no software R, e o fluxo completo do procedimento é descrito, a seguir, no Algoritmo 2.

Algorithm 2: Algoritmo BJ-ELM-*t-Student* Adaptativo com Boosting L2

Data: Dados de treino (X, y),
número de neurônios ocultos Q,
número de aprendizes base M,
taxa de aprendizado γ,
número pequeno positivo φ

Result: Modelo final $\widehat{f}(\mathbf{x})$

Inicializar $y^{*(0)} = y$, $k = 0$, $P^{(0)} = I$, $f(x)^{(0)} = 0,001$, **função de ativação inicial** $g(.) = I$;

repeat

$k \leftarrow k + 1$;

`// Etapa de Boosting L2 com ELM`

Ajustar o logaritmo do tempo de sobrevivência utilizando um modelo de boosting baseado em ELM ponderado por matriz P:

$$f(\mathbf{x}_i)^{(k)} = \sum_{m=1}^{M} G(\mathbf{x}_i^{\top}\mathbf{W}^{(k,m)} + \mathbf{B}^{(k,m)})\boldsymbol{\alpha}^{(k,m)}, \quad i = 1, \ldots, n$$

Calcular os resíduos atuais: $\epsilon_i^{(k)} = y_i - \widehat{f}(\mathbf{x}_i)^{(k)}, \quad i = 1, \ldots, n$;

Atualizar $P^{(k)}$;

Ordenar $\epsilon_1 < \epsilon_2 < ... < \epsilon_n$ e reorganizar os dados de sobrevivência imputados $(y_i,\delta_i,\mathbf{x}_i)$;

Atualizar $y_i^{*(k)} = \delta_i y_i + (1 - \delta_i)\left[\widehat{f}(\mathbf{x}_i)^{(k)} + \widehat{S}(\epsilon_i^{(k)})^{-1}\sum_{\epsilon_j^{(k)} > \epsilon_i^{(k)}} \epsilon_j^{(k)}\delta_j \boldsymbol{\Delta}\widehat{S}(\epsilon_j^{(k)})\right]$;

`// Etapas adaptativas realizadas ao final de cada ciclo` k

Aplicar o teste RESET sobre $y^{*(k)}$ para decidir a função de ativação:

- Se linearidade confirmada $\rightarrow$ manter $g(.) = I$;
- Caso contrário $\rightarrow$ atualizar $g(.) =$ sigmoide.

Atualizar o grau de liberdade ν da distribuição *t-Student* via otimização:

$$\nu^{(k)} = \arg\max_{\nu} \sum_i \log \mathbb{K}_i, \quad \mathbb{K}_i = \frac{2^{\frac{1}{2}}\nu^{\frac{\nu}{2}}\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(d_i + \nu)^{\frac{\nu+1}{2}}}$$

until $\frac{\left\|\widehat{f}(\mathbf{x})^{(k)} - \widehat{f}(\mathbf{x})^{(k-1)}\right\|}{\left\|\widehat{f}(\mathbf{x})^{(k)}\right\|} < \varphi$;

Retornar $\widehat{f}(\mathbf{x})$;

4 RESULTADOS

Neste capítulo são apresentados os resultados de estudos destinados a avaliar e validar o desenvolvimento de um modelo adaptativo para análise de sobrevivência. As avaliações utilizam duas métricas amplamente empregadas na literatura, o **C-Index** e o **IBS**, que permitem mensurar, respectivamente, a capacidade discriminativa e a calibração dos modelos.

Os estudos seguem a abordagem proposta por Kong et al. [17], contemplando duas etapas principais. A primeira consiste na aplicação de diferentes algoritmos sobre diversos conjuntos de dados simulados, criados para avaliar separadamente características fundamentais do modelo adaptativo. A segunda etapa envolve a análise de seis conjuntos de dados clínicos reais, amplamente utilizados na literatura e disponíveis em pacotes do software R [46].

4.1 MEDIDAS DE DESEMPENHO

4.1.1 Concordance Index (C-Index)

O C-index foi proposto por Harrell et al. (1996) [47]. É uma medida frequentemente usada para avaliar a precisão preditiva de modelos em análise de sobrevivência. Reflete a consistência entre a previsão de sobrevivência de um modelo e a situação real de sobrevivência, medindo a proporção dos pares de indivíduos cujas previsões de sobrevivência têm o mesmo ordenamento com seu tempo de evento verdadeiro e todos os pares de indivíduos, para os quais o tempo de evento é comparável. Esta medida pode ser calculada por

$$C - Index(f) = \frac{\sum_{i=1}^{n} \sum_{j \neq i} I((f(x_i) - f(x_j)(y_i - y_j) \geq 0)}{\sum_{i=1}^{n} \sum_{j \neq i} comp(i,j)} \tag{4.1}$$

onde f representa um modelo de previsão de sobrevivência e $comp(i,j)$ representa um par de indivíduos cujo tempo de sobrevivência real é comparável e pode ser calculado como

$$comp(i,j) = \begin{cases} 1 & \text{se} \quad (\delta_i = \delta_j = 1)\ ou\ (\delta_i = 1, \delta_j = 0\ e\ y_i \leq y_j) \\ 0 & \text{outro caso .} \end{cases} \tag{4.2}$$

Na análise de sobrevivência, o valor do C-Index está entre 0 e 1. Um valor maior do índice indica uma consistência mais alta entre resultados de predição do modelo e o tempo real de sobrevivência, e assim, um melhor desempenho preditivo.

4.1.2 Integrated Brier Score (IBS)

O Brier score (BS), um erro preditivo dependente do tempo na Análise de Sobrevivência, proposto por Graf et al. (1999) [48], é definido como a média dos quadrados das diferenças entre a probabilidade de sobrevivência prevista e o estado de sobrevivência real,

ponderado pela probabilidade inversa de censura. A medida BS no tempo t é calculada como

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(\widehat{S}(t|x_i))^2}{\widehat{G}(y_i)}.I(y_i < t, \delta_i = 1) + \frac{(1-\widehat{S}(t|x_i))^2}{\widehat{G}(y_i)}.I(y_i \geq t)\right], \tag{4.3}$$

onde $\widehat{S}(.)$ é a função de sobrevivência prevista pelo modelo e $\widehat{G}(.)$ é o estimador KM da distribuição de censura.

De (4.3), pode-se observar que o BS é um erro quadrático de predição que depende do ponto no tempo t. A seleção de ponto temporal t pode levar a grandes diferenças na avaliação do desempenho preditivo do modelo. Assim, uma medida de erro de previsão mais abrangente é o denominado IBS, que é definido como a forma integral do BS e não depende da seleção de um único ponto temporal t. O valor do IBS é calculado por

$$IBS = \frac{1}{max_i(y_i)}\int_0^{max_i(y_i)} BS(t)dt. \tag{4.4}$$

Na análise de sobrevivência, quanto menor for o IBS, melhor será o desempenho do modelo de sobrevivência.

4.2 ESTUDOS DE SIMULAÇÃO

Nos estudos de simulação são explorados diferentes cenários relacionados às características do conjunto de dados. São aplicados dois modelos existentes e três variações do modelo proposto, com o objetivo de avaliar a robustez no sentido de verificar a capacidade dos modelos em lidar com dados extremos, bem como o desempenho em termos de capacidade discriminativa (C-Index) e calibração (IBS) em comparação com modelos já consolidados na literatura. Para cada cenário foram simuladas 30 amostras Monte Carlo, e em cada uma delas aplicaram-se os cinco modelos avaliados. A seguir, apresentam-se em detalhe os cenários considerados e a descrição dos modelos testados.

4.2.1 Especificação dos Cenários Simulados

Foram considerados seis cenários para o conjunto de dados. As configurações destes cenários são descritas a seguir.

- **Cenário 1: Dados com efeitos lineares e não correlacionados**

  O tempo do evento segue um modelo AFT log-normal. O logaritmo do tempo do evento T é gerado a partir do modelo

  $$T = 0.5 + \mathbf{X}^\top\boldsymbol{\beta} + \epsilon,$$

  onde

O termo de erro aleatório ϵ é gerado a partir de $N(0, 3)$.

Os elementos de $\boldsymbol{\beta}$ são $0.4, 0, 5, 0.6, 0.7, 0.8$ repetidos 6 vezes.

$\mathbf{X}$ é um vetor de covariáveis de dimensão 30 extraído de $\mathbf{N}_{30}(\mathbf{0}, \boldsymbol{\Sigma})$ onde $\boldsymbol{\Sigma}$ é matriz diagonal de dimensão 30 com valores de diagonal 0.95 e fora da diagonal 0.

O tempo de censura é gerado a partir de uma distribuição uniforme $[0, 20]$.

- **Cenário 2: Dados com efeitos lineares e interações**

  O tempo do evento segue um modelo AFT log-normal. O logaritmo do tempo do evento T é gerado a partir do modelo

  $$T = 0.5 + \mathbf{X}^\top \boldsymbol{\beta} + \epsilon,$$

  onde

  O termo de erro aleatório ϵ é gerado a partir de $N(0, 3)$.

  Os primeiros 15 elementos de $\boldsymbol{\beta}$ são 0.4 e os últimos 15 elementos são 0.

  $\mathbf{X}$ é um vetor de covariáveis de dimensão 30 extraído de $\mathbf{N}_{30}(\mathbf{0}, \boldsymbol{\Sigma})$ onde $\boldsymbol{\Sigma} = diag([\mathbf{J}, \mathbf{J}, \mathbf{J}, \mathbf{I}])$.

  $\mathbf{J}$ é matriz diagonal de dimensão 5 com valores de diagonal $1, 01$ e fora da diagonal 1, e $\mathbf{I}$ é uma matriz identidade de ordem 15.

  O tempo de censura é gerado a partir de uma distribuição uniforme $[0, 20]$.

  Esta configuração corresponde ao primeiro cenário de dados com efeitos lineares apresentado por Wang et al. (2010) [50].

- **Cenário 3: Dados com efeitos não lineares**

  O logaritmo do tempo do evento T é gerado a partir do modelo

  $$T = f(\mathbf{X}^\top) + \epsilon$$

  onde

  O termo de erro aleatório ϵ é gerado a partir de $N(0, 0.75)$.

  $\mathbf{X}$ é um vetor de covariáveis de dimensão 4 gerado de uma distribuição normal com média zero e uma estrutura autoregresiva de correlação, tal que $corr(x_i, x_j) = 0.7^{|i-j|}, i, j = 1, 2, 3, 4$.

  O logaritmo do tempo de censura C é gerado a partir de uma distribuição $N(0, 0.75)$.

  Segundo Wang et al. (2010) [50], $f(\mathbf{X}) = f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4)$, onde $f_1(X_1) = 4X_1^2)$,$f_2(X_2) = sin(6X_2^2)$, $f_3(X_3) = cos(6X_3) - 1$, e $f_4(X_4) = 4X_4^3) + X_4^2$.

- **Cenário 4: Dados com efeitos lineares e não correlacionados com erros t-Student**

  Mantém todas as configurações do cenário 1, menos a geração do erro aleatório que agora é gerado por uma distribuição *t-Student* com 10 graus de liberdade.

- **Cenário 5: Dados com efeitos lineares e interações com erros t-Student**

  Mantém todas as configurações do cenário 2, menos a geração do erro aleatório que agora é gerado por uma distribuição *t-Student* com 10 graus de liberdade.

- **Cenário 6: Dados com efeitos não lineares com erros t-Student**

  Mantém todas as configurações do cenário 3, menos a geração do erro aleatório que agora é gerado por uma distribuição *t-Student* com 10 graus de liberdade.

4.2.2 Modelos avaliados

Foram avaliados cinco modelos, dois dos quais são conhecidos na literatura e três resultantes das variações de algoritmo proposto. A descrição destes modelos se apresentam a seguir:

- **Modelo CPH:** Corresponde ao modelo de Riscos Proporcionais de Cox [18] , é o mais conhecido e usado na Análise de Sobrevivência.
- **Modelo BJ-ELM:** Proposto por Kong e Zhan (2023) [17]. Corresponde ao Modelo de boosting de Buckley–James baseado em ELM. Este modelo é mais eficiente que outros modelos mais conhecidos quando existe não linearidade.
- **Modelo BJ-ELML:** Primeira variação da proposta. Considera a função de ativação $g(x) = x$.
- **Modelo BJ-ELMR:** Proposta robusta que considera distribuição *t-Student* com 10 graus de liberdade e função de ativação $g(x)$ sigmoide.
- **Modelo BJ-ELMLR:** Proposta robusta que considera distribuição *t-Student* com 10 graus de liberdade e função de ativação $g(x) = x$.

4.2.3 Avaliação do cenário 1

O cenário 1 simula uma situação considerada ideal para a aplicação de modelos lineares, caracterizada pela independência entre covariáveis (correlação praticamente nula) e erros seguindo uma distribuição normal. Nessas condições, a estrutura linear é suficiente para representar adequadamente os dados. A Tabela 3 mostra que, para ambas as medidas de desempenho, os piores resultados correspondem aos modelos não lineares. Entre os modelos do tipo BJ, o BJ-ELML apresenta melhor desempenho segundo ambas métricas. O

modelo CPH apresenta desempenho muito similar aos modelos BJ lineares, o que pode ser explicado pela condição ideal de linearidade do dado simulado. A Figura 4, que apresenta em conjunto os box-plots das distribuições de C-Index (na parte superior) e de IBS (na parte inferior) para os cinco métodos avaliados, indica variabilidade equivalente entre eles no C-Index e reforça a conclusão anterior sobre o melhor desempenho dos métodos lineares. Em relação ao IBS, observa-se que o modelo CPH apresenta maior variabilidade quando comparado aos demais quatro métodos, confirmando a tendência já identificada nas medidas de desempenho.

Tabela 3 – Avaliação de Cenário 1: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
CPH	0.756 (0.727,0.785)	0.180 (0.152,0.208)
BJ-ELM	0.750 (0.717,0.782)	0.188 (0.174,0.202)
BJ-ELML	0.760 (0.727,0.793)	0.181 (0.167,0.195)
BJ-ELMR	0.751 (0.718,0.784)	0.188 (0.173,0.203)
BJ-ELMLR	0.756 (0.726,0.787)	0.183 (0.169,0.198)

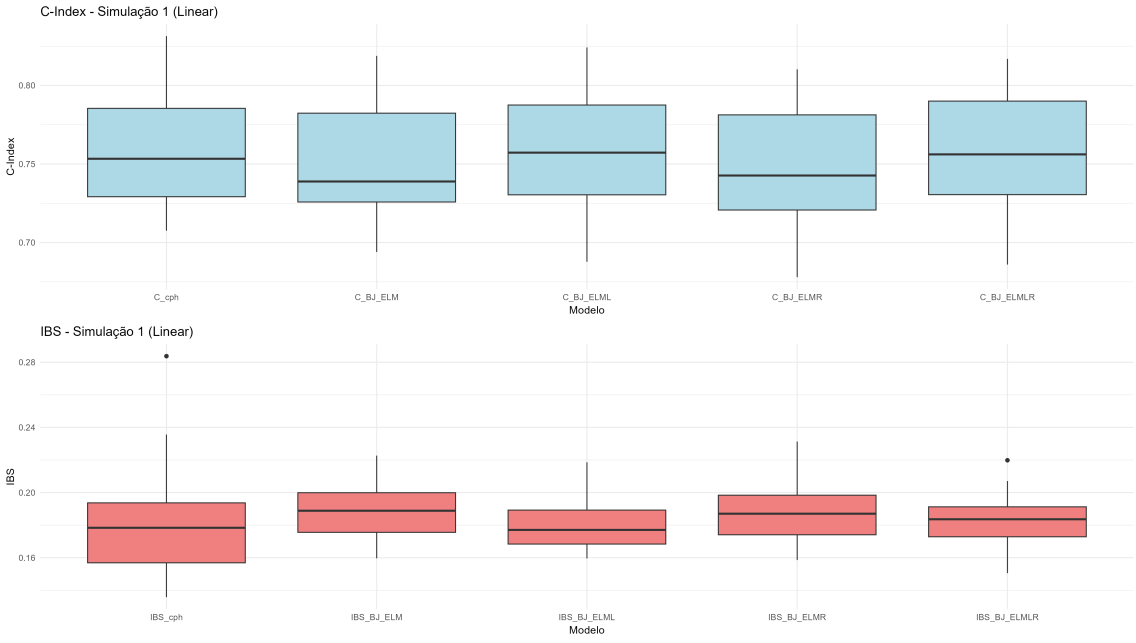

Figura 4 – Avaliação de Cenário 1: Box-Plot para C-Index e IBS em modelos avaliados.

4.2.4 Avaliação do cenário 2

O cenário 2 simula uma situação em que há interação entre algumas covariáveis. A Tabela 4 mostra um equilíbrio nos desempenhos de todos os modelos avaliados. A Figura 5, que apresenta em conjunto os box-plots das distribuições de C-Index (na parte superior) e de IBS (na parte inferior) para os cinco métodos, indica menor variabilidade no modelo BJ-ELMR e reforça a conclusão de similaridade no desempenho entre os métodos. Em relação ao IBS, observa-se novamente que o modelo CPH apresenta maior variabilidade quando comparado aos demais quatro modelos.

Tabela 4 – Avaliação de Cenário 2: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
CPH	0.753 (0.713,0.792)	0.176 (0.148,0.203)
BJ-ELM	0.759 (0.719,0.799)	0.180 (0.165,0.194)
BJ-ELML	0.758 (0.721,0.796)	0.178 (0.162,0.194)
BJ-ELMR	0.759 (0.719,0.798)	0.180 (0.164,0.195)
BJ-ELMLR	0.757 (0.720,0.793)	0.178 (0.161,0.195)

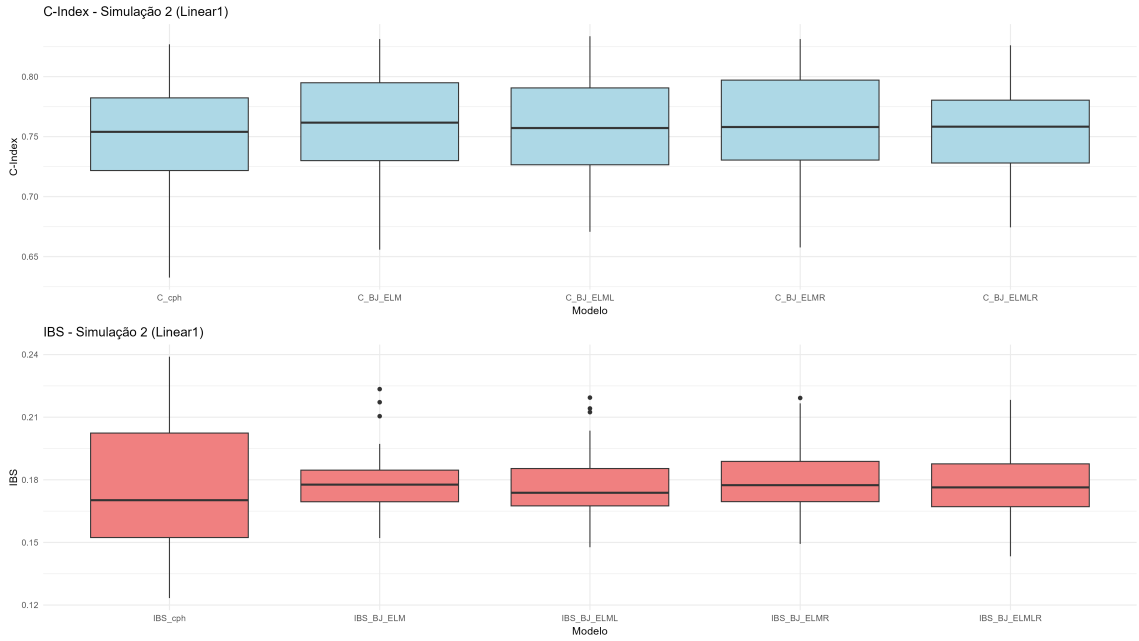

Figura 5 – Avaliação de Cenário 2: Box-Plot para C-Index e IBS em modelos avaliados.

4.2.5 Avaliação do cenário 3

O cenário 3 simula uma situação em que há não linearidade entre as covariáveis. A Tabela 5 evidencia o melhor desempenho dos modelos não lineares em ambas as medidas de avaliação, diferença que se mostra bastante clara. A Figura 6, que apresenta em conjunto os box-plots das distribuições de C-Index (na parte superior) e de IBS (na parte inferior) para os cinco métodos, confirma essa conclusão ao mostrar menor variabilidade nos modelos não lineares. Em relação ao IBS, observa-se que todas as medidas apresentam valores elevados em algumas repetições, reforçando a complexidade deste cenário e a necessidade de modelos capazes de lidar com estruturas não lineares.

Tabela 5 – Avaliação de Cenário 3: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
CPH	0.882 (0.838,0.927)	0.232 (0.000,0.467)
BJ-ELM	0.917 (0.879,0.955)	0.185 (0.038,0.333)
BJ-ELML	0.878 (0.834,0.923)	0.213 (0.033,0.394)
BJ-ELMR	0.921 (0.884,0.958)	0.181 (0.032,0.330)
BJ-ELMLR	0.880 (0.836,0.924)	0.226 (0.016,0.437)

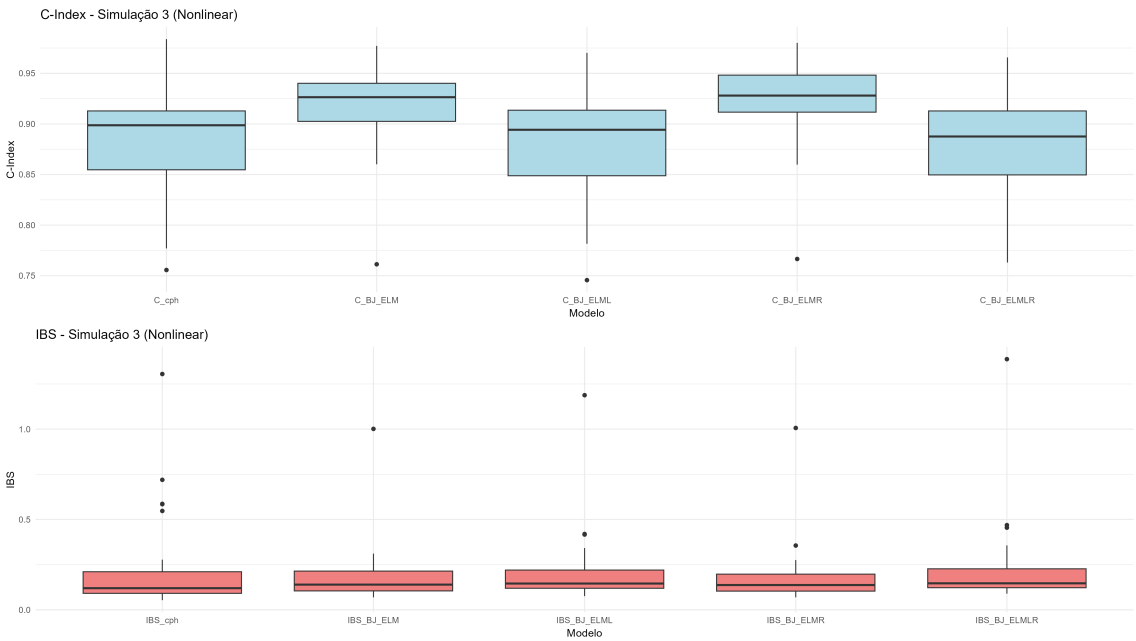

Figura 6 – Avaliação de Cenário 3: Box-Plot para C-Index e IBS em modelos avaliados.

4.2.6 Avaliação do cenário 4

O cenário 4 reproduz o cenário 1, porém com a inclusão de dados extremos, e apresenta avaliações semelhantes. A Tabela 6 mostra que os piores desempenhos correspondem aos modelos não lineares, enquanto o modelo CPH apresenta resultados muito próximos aos modelos BJ lineares. A Figura 7, que reúne os box-plots das distribuições de C-Index (na parte superior) e de IBS (na parte inferior) para os cinco métodos, confirma o melhor desempenho dos modelos lineares, com menor variabilidade no C-Index. Em relação ao IBS, observa-se que o CPH apresenta valores médios mais baixos, mas também maior variabilidade quando comparado aos demais quatro modelos.

Tabela 6 – Avaliação de Cenário 4: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
CPH	0.901 (0.884,0.918)	0.080 (0.058,0.103)
BJ-ELM	0.883 (0.859,0.907)	0.142 (0.132,0.153)
BJ-ELML	0.896 (0.877,0.915)	0.136 (0.125,0.146)
BJ-ELMR	0.886 (0.862,0.909)	0.142 (0.131,0.153)
BJ-ELMLR	0.897 (0.876,0.917)	0.135 (0.124,0.146)

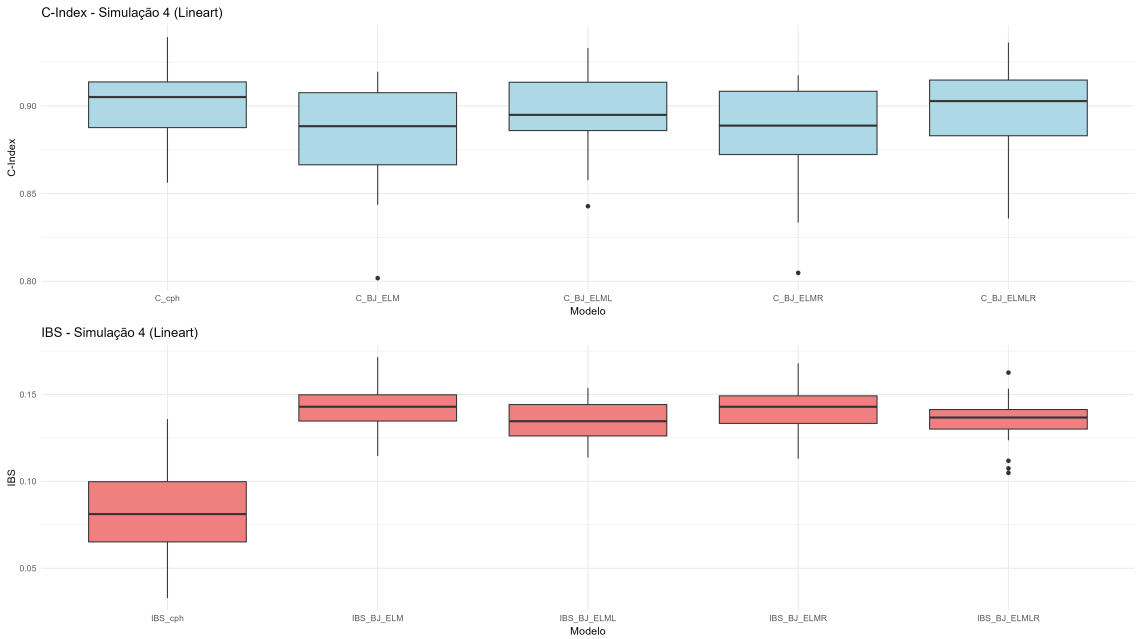

Figura 7 – Avaliação de Cenário 4: Box-Plot para C-Index e IBS em modelos avaliados.

4.2.7 Avaliação do cenário 5

O cenário 5 simula uma situação em que há interação entre algumas covariáveis combinada com a presença de dados extremos. A Tabela 7 evidencia um equilíbrio nos desempenhos de todos os modelos avaliados. A Figura 8, que apresenta em conjunto os box-plots das distribuições de C-Index (na parte superior) e de IBS (na parte inferior), confirma essa tendência ao mostrar desempenhos semelhantes entre os métodos. Em relação ao IBS, observa-se que o modelo CPH apresenta valores médios mais favoráveis, embora com maior variabilidade em comparação aos demais quatro modelos. Os resultados guardam estreita semelhança com aqueles obtidos para o cenário 2.

Tabela 7 – Avaliação de Cenário 5: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
CPH	0.908 (0.888,0.928)	0.072 (0.049,0.094)
BJ-ELM	0.910 (0.890,0.930)	0.128 (0.116,0.141)
BJ-ELML	0.906 (0.885,0.927)	0.131 (0.117,0.145)
BJ-ELMR	0.909 (0.891,0.928)	0.131 (0.120,0.142)
BJ-ELMLR	0.908 (0.888,0.928)	0.130 (0.116,0.145)

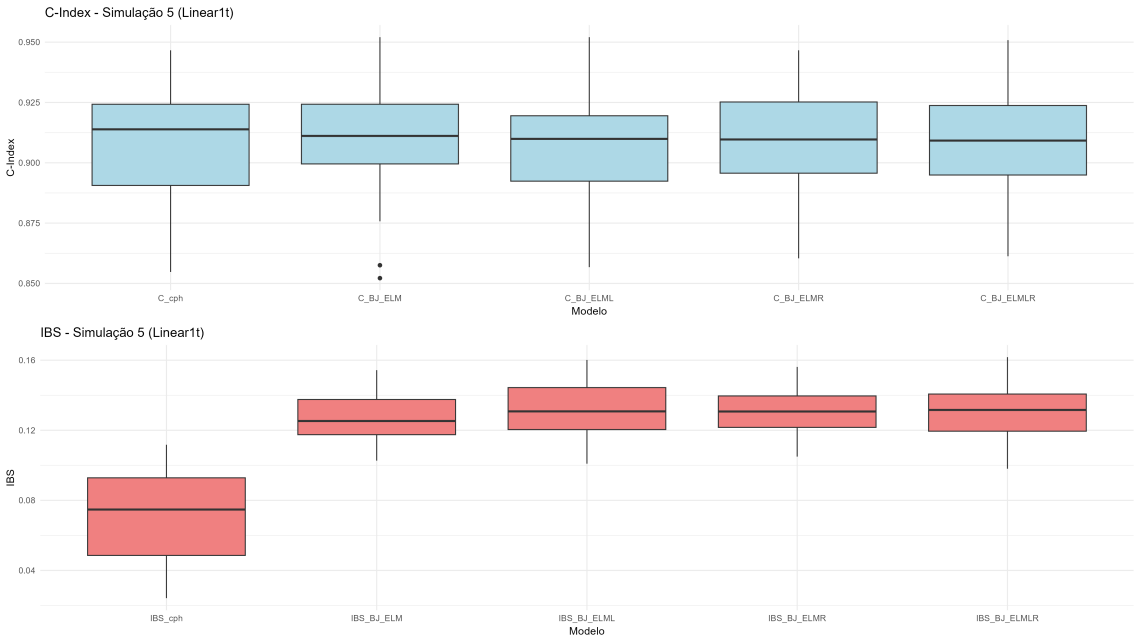

Figura 8 – Avaliação de Cenário 5: Box-Plot para C-Index e IBS em modelos avaliados.

4.2.8 Avaliação do cenário 6

O cenário 6 simula uma situação em que há não linearidade entre as covariáveis combinada com a presença de dados extremos. A Tabela 8 evidencia o melhor desempenho dos modelos não lineares, destacando-se entre eles o BJ-ELMR. A Figura 9, que apresenta em conjunto os box-plots das distribuições de C-Index (na parte superior) e de IBS (na parte inferior), confirma essa conclusão ao mostrar menor variabilidade nos modelos não lineares. De modo geral, as avaliações obtidas neste cenário são similares às verificadas no cenário 3.

Tabela 8 – Avaliação de Cenário 6: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
CPH	0.883 (0.840,0.926)	0.194 (0.039,0.348)
BJ-ELM	0.911 (0.871,0.952)	0.153 (0.088,0.218)
BJ-ELML	0.873 (0.826,0.920)	0.176 (0.094,0.259)
BJ-ELMR	0.914 (0.876,0.952)	0.148 (0.096,0.201)
BJ-ELMLR	0.878 (0.832,0.925)	0.176 (0.101,0.251)

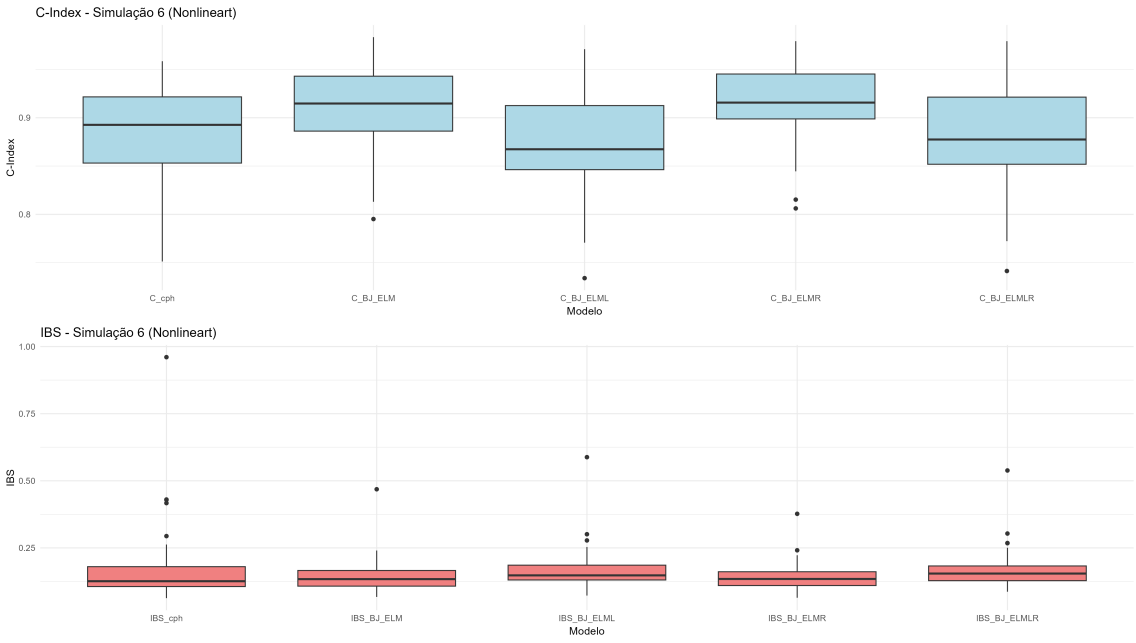

Figura 9 – Avaliação de Cenário 6: Box-Plot para C-Index e IBS em modelos avaliados.

4.3 APLICAÇÃO EM DADOS REAIS

Neste seção apresentam-se dos conjuntos de aplicações sobre seis conjuntos de dados reais presentes em diversas pacotes do software R [46]. Sobre estes conjuntos de dados serão feitos dois estudos. O primeiro estudo avalia a inclusão de robustez no modelo proposto por Kong et al. (2023) [17]. O segundo estudo avalia a performance do algoritmo adaptativo proposto, que já inclui a opção de linearidade ou não e o uso da distribuição *t-Student* para garantir robustez ao respeito de dados extremos.

A descrição dos seis conjuntos de dados são apresentados na Tabela 9 a seguir:

Tabela 9 – Descrição do bancos de dados usados.

Nome	Origem	Casos	Censura (%)	Número de covariáveis
Pbc	randomForestSRC	418	61%	17
Lung	survival	226	2%	8
WPBC	TH.data	198	24%	32
StageC	rpart	146	63%	6
Veteran	survival	137	7%	6
Prca	SubgrPlots	475	29%	7

4.3.1 Avaliação de inclusão de robustez no modelo

Nesta subseção apresenta-se os resultados da aplicação do modelo proposto em dados reais, e avalia-se o seu desempenho em contextos práticos. Antes de iniciar esta avaliação devemos resumir algumas conclusões gerais obtidas dos cenários simulados: Primeiro, o método CPH é adequado para algumas situações, entretanto apresenta variabilidade maior que os modelos BJ, o que faz que seja menos consistente. Por outro lado, os modelos BJ acomodam-se melhor as características de linearidade ou não dos conjuntos simulados e são mais consistentes, assim, o uso de uma adequada função de ativação torna-se importante. Finalmente, existe um pequeno indicativo da necessidade de uso de métodos robustos, a aplicação do método robusto sobre dados reais visa clarificar a necessidade do uso de este tipo de modelo.

Os seis conjuntos de dados nos quais se avaliam os métodos robustos propostos foram analisados por Kong et al. (2023) [17], eles concluem que, para todos eles, o método BJ-ELM é adequado. Considerando esta conclusão, o modelos BJ-ELMR com distintos graus de liberdade serão avaliados e comparados com os resultados do BJ-ELM, usando as medidas de desempenho C-Index e IBS.

Como o método BJ-ELMR proposto depende da escolha do grau de liberdade da distribuição *t-Student*, uma grade de valores é definida. Assim, são avaliados modelos BJ-ELMR com valores $(5, 10, 15, 20, 25, 30)$ no grau de liberdade.

Observe-se as Tabelas 10-15 onde se observa que cada conjunto de dados obtém um melhor desempenho em valores diferentes de graus de liberdade, assim, mostra-se a importância do uso dos graus de liberdade para melhorar os desempenhos de predição. Quando o grau de liberdade é muito grande os resultados se aproximam do BJ-ELM.

Tabela 10 – Pbc: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
BJ-ELM	0.829 (0.816,0.842)	0.185 (0.155,0.215)
BJ-ELMR (5)	0.829 (0.815,0.843)	0.183 (0.153,0.212)
BJ-ELMR (10)	0.827 (0.814,0.840)	0.185 (0.156,0.215)
BJ-ELMR (15)	0.825 (0.812,0.839)	0.185 (0.155,0.215)
BJ-ELMR (20)	0.830 (0.817,0.843)	0.183 (0.153,0.214)
BJ-ELMR (25)	0.828 (0.815,0.841)	0.183 (0.154,0.211)
BJ-ELMR (30)	0.826 (0.814,0.838)	0.186 (0.155,0.218)

Tabela 11 – Lung: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
BJ-ELM	0.617 (0.600,0.634)	0.207 (0.174,0.240)
BJ-ELMR (5)	0.627 (0.608,0.646)	0.204 (0.172,0.235)
BJ-ELMR (10)	0.626 (0.610,0.642)	0.204 (0.172,0.237)
BJ-ELMR (15)	0.622 (0.605,0.638)	0.206 (0.172,0.239)
BJ-ELMR (20)	0.621 (0.603,0.639)	0.207 (0.173,0.241)
BJ-ELMR (25)	0.623 (0.606,0.641)	0.205 (0.173,0.237)
BJ-ELMR (30)	0.624 (0.606,0.642)	0.206 (0.173,0.239)

Tabela 12 – WPBC: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
BJ-ELM	0.643 (0.622,0.663)	0.160 (0.154,0.165)
BJ-ELMR (5)	0.649 (0.629,0.669)	0.159 (0.154,0.164)
BJ-ELMR (10)	0.623 (0.599,0.648)	0.163 (0.158,0.168)
BJ-ELMR (15)	0.648 (0.623,0.674)	0.159 (0.154,0.164)
BJ-ELMR (20)	0.644 (0.621,0.667)	0.159 (0.154,0.163)
BJ-ELMR (25)	0.642 (0.626,0.659)	0.159 (0.153,0.166)
BJ-ELMR (30)	0.633 (0.614,0.653)	0.161 (0.155,0.166)

Tabela 13 – StageC: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
BJ-ELM	0.724 (0.686,0.762)	0.167 (0.150,0.185)
BJ-ELMR (5)	0.726 (0.694,0.757)	0.170 (0.151,0.190)
BJ-ELMR (10)	0.717 (0.683,0.751)	0.167 (0.150,0.183)
BJ-ELMR (15)	0.719 (0.685,0.754)	0.168 (0.150,0.185)
BJ-ELMR (20)	0.722 (0.684,0.759)	0.168 (0.151,0.185)
BJ-ELMR (25)	0.721 (0.686,0.757)	0.169 (0.151,0.186)
BJ-ELMR (30)	0.722 (0.687,0.757)	0.170 (0.149,0.192)

Tabela 14 – Veteran: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
BJ-ELM	0.761 (0.740,0.783)	0.423 (0.159,0.687)
BJ-ELMR (5)	0.757 (0.736,0.779)	0.420 (0.160,0.681)
BJ-ELMR (10)	0.755 (0.732,0.777)	0.416 (0.153,0.679)
BJ-ELMR (15)	0.751 (0.727,0.776)	0.426 (0.158,0.694)
BJ-ELMR (20)	0.753 (0.733,0.774)	0.412 (0.160,0.664)
BJ-ELMR (25)	0.756 (0.734,0.778)	0.430 (0.158,0.701)
BJ-ELMR (30)	0.751 (0.728,0.773)	0.427 (0.162,0.692)

Tabela 15 – Prca: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
BJ-ELM	0.638 (0.625,0.651)	0.400 (0.337,0.464)
BJ-ELMR (5)	0.641 (0.628,0.654)	0.403 (0.340,0.465)
BJ-ELMR (10)	0.640 (0.628,0.652)	0.402 (0.339,0.464)
BJ-ELMR (15)	0.638 (0.625,0.651)	0.400 (0.339,0.461)
BJ-ELMR (20)	0.640 (0.628,0.653)	0.401 (0.338,0.464)
BJ-ELMR (25)	0.640 (0.626,0.653)	0.401 (0.339,0.464)
BJ-ELMR (30)	0.639 (0.627,0.651)	0.402 (0.339,0.465)

4.3.2 Avaliação do modelo adaptativo proposto

Nesta subseção apresentam-se os resultados da aplicação de cinco modelos sobre os seis conjuntos reis usados encontrados na literatura. O desempenho foi avaliado por meio das métricas **C-Index** e **IBS**, amplamente utilizadas em análises de sobrevivência.

Antes da apresentação e discussão de resultados, apresenta-se a descrição do procedimento usado para estas avaliações. O procedimento seguido é diferente do usado na primeira avaliação. Desta vez, considera-se que os modelos podem ser lineares ou não de forma que a proposta adaptativa possa ser avaliada completamente.

O procedimento baseia-se na validação cruzada, que será repetida 20 vezes. Em

cada uma das repetições, os dados foram particionados em 80% para treinamento e 20% para teste e os modelos avaliados foram ajustados no conjunto de treinamento e avaliados no conjunto de teste, gerando valores de C-Index e IBS. A repetição desse processo permitiu capturar a variabilidade dos resultados e, a partir das execuções, construir tabelas e gráficos que sintetizam o desempenho comparativo dos modelos.

Os modelos comparados nesta avaliação são:

- **CPH (Cox Proportional Hazards)**: modelo clássico e frequentemente usado em estudos de sobrevivência. Baseia-se em relações lineares entre covariáveis e risco.
- **BJ-LS (Buckley-James)**: modelo linear proposto como alternativa ao modelo de Cox.
- **BJ-ELM (Extreme Learning Machine)**: modelo proposto por Kong et al. [17], que considera não linearidade e usa ELM.
- **BJ-ELML (Adaptativo)**: versão adaptativa que avalia linearidade a partir do teste RESET, e usa funções de ativação identidade ou sigmoide no proceso ELM, é uma primeira variação de BJ-ELM.
- **BJ-ELMLR (Adaptativo Robusto)**: versão adaptativa baseada no BJ-ELML que inclue o tratamento de dados extremos com o uso da distribuição *t-Student*.

Um primeiro resultado relevante associado ao modelo proposto BJ-ELMLR (Adaptativo Robusto) refere-se aos valores médios dos graus de liberdade da distribuição *t-Student*, que variam de acordo com cada conjunto de dados analisado. Especificamente, os valores estimados (arredondados) foram 26, 16, 19, 30, 17 e 27 para os bancos Pbc, Lung, WPBC, StageC, Veteran e Prca, respectivamente. Esses resultados evidenciam que o algoritmo proposto incorpora, de forma adaptativa, a informação proveniente de observações extremas quando necessário, ajustando a robustez do modelo às características específicas de cada base de dados.

Comparativamente, alguns resultados diferem daqueles reportados por Kong et al. (2023) [17]. Apenas dois conjuntos de dados (Pbc e Lung) apresentaram evidências de não linearidade segundo a avaliação pelo C-Index, enquanto os demais mostraram melhor desempenho sob modelos lineares. Essa divergência pode ser atribuída ao procedimento de validação cruzada adotado: enquanto Kong et al. [17] realizaram seleções independentes de dados para teste e validação em cada método avaliado, neste estudo foi utilizada uma única partição de dados aplicada de forma uniforme aos cinco métodos comparados. Ainda assim, os resultados demonstram que os métodos adaptativos propostos acompanham as variações estruturais dos dados, confirmando a flexibilidade da abordagem.

As Tabelas 16–21 evidenciam que os modelos adaptativos baseados em BJ-ELM acompanham o desempenho dos modelos tradicionais. Em cenários com indicativos de

linearidade, os modelos CPH e BJ-LS alcançam valores superiores de C-Index e IBS, sendo que os modelos adaptativos mantêm desempenho próximo. Por outro lado, em situações com indicativos de não linearidade, os modelos da classe BJ-ELM apresentam métricas de desempenho mais favoráveis. Em determinados casos, os métodos adaptativos não atingem os melhores valores absolutos, mas permanecem próximos dos modelos considerados mais adequados, reforçando sua capacidade de adaptação às diferentes estruturas dos dados.

As Figuras 10–15 apresentam os box-plots das distribuições de C-Index e IBS obtidas em 20 repetições para os modelos avaliados, cada gráfico correspondendo a um conjunto de dados específico.

A análise dos gráficos evidencia que os modelos adaptativos da classe BJ-ELM apresentam desempenhos próximos ou superiores aos métodos tradicionais não adaptativos (CPH, BJ-LS ou BJ-ELM). Esses resultados são consistentes com aqueles apresentados nas tabelas, indicando que modelos adaptativos baseados em BJ-ELM podem ser considerados uma alternativa flexível, capaz de acompanhar a estrutura e as características dos dados em comparação com abordagens tradicionais consolidadas.

Tabela 16 – Pbc: Resultados comparativos para C-Index e IBS nos modelos avaliados.

Modelo	C-Index	IBS
CPH	0.804 (0.788, 0.819)	0.193 (0.140, 0.245)
BJ-LS	0.816 (0.800, 0.832)	0.219 (0.165, 0.272)
BJ-ELM	0.824 (0.809, 0.839)	0.184 (0.153, 0.214)
BJ-ELML (Adaptativo)	0.820 (0.806, 0.835)	0.184 (0.154, 0.215)
BJ-ELMLR (Adaptativo)	0.820 (0.805, 0.834)	0.183 (0.154, 0.213)

Tabela 17 – Lung: Resultados comparativos para C-Index e IBS nos modelos avaliados.

Modelo	C-Index	IBS
CPH	0.590 (0.572, 0.608)	0.210 (0.191, 0.229)
BJ-LS	0.578 (0.558, 0.598)	0.200 (0.180, 0.220)
BJ-ELM	0.596 (0.572, 0.619)	0.196 (0.178, 0.214)
BJ-ELML (Adaptativo)	0.597 (0.573, 0.621)	0.196 (0.178, 0.214)
BJ-ELMLR (Adaptativo)	0.601 (0.577, 0.625)	0.195 (0.177, 0.214)

Tabela 18 – WPBC: Resultados comparativos para C-Index e IBS nos modelos avaliados.

Modelo	C-Index	IBS
CPH	0.655 (0.628, 0.682)	0.169 (0.156, 0.182)
BJ-LS	0.648 (0.619, 0.677)	0.169 (0.155, 0.176)
BJ-ELM	0.652 (0.622, 0.682)	0.163 (0.154, 0.171)
BJ-ELML (Adaptativo)	0.651 (0.626, 0.676)	0.165 (0.156, 0.174)
BJ-ELMLR (Adaptativo)	0.652 (0.628, 0.676)	0.165 (0.157, 0.174)

Tabela 19 – StageC: Resultados comparativos para C-Index e IBS nos modelos avaliados.

Modelo	C-Index	IBS
CPH	0.731 (0.702, 0.760)	0.198 (0.163, 0.233)
BJ-LS	0.697 (0.666, 0.727)	0.178 (0.157, 0.198)
BJ-ELM	0.702 (0.677, 0.727)	0.191 (0.163, 0.219)
BJ-ELML (Adaptativo)	0.736 (0.711, 0.760)	0.188 (0.160, 0.216)
BJ-ELMLR (Adaptativo)	0.727 (0.703, 0.751)	0.192 (0.163, 0.220)

Tabela 20 – Veteran: Resultados comparativos para C-Index e IBS nos modelos avaliados.

Modelo	C-Index	IBS
CPH	0.697 (0.677, 0.717)	0.113 (0.028, 0.197)
BJ-LS	0.713 (0.694, 0.731)	0.122 (0.016, 0.228)
BJ-ELM	0.687 (0.663, 0.710)	0.120 (0.042, 0.199)
BJ-ELML (Adaptativo)	0.716 (0.696, 0.736)	0.119 (0.039, 0.199)
BJ-ELMLR (Adaptativo)	0.717 (0.697, 0.737)	0.116 (0.039, 0.194)

Tabela 21 – Prca: Resultados comparativos para C-Index e IBS nos modelos avaliados.

Modelo	C-Index	IBS
CPH	0.621 (0.608, 0.633)	0.422 (0.338, 0.505)
BJ-LS	0.606 (0.595, 0.617)	0.431 (0.348, 0.513)
BJ-ELM	0.617 (0.606, 0.629)	0.415 (0.340, 0.490)
BJ-ELML (Adaptativo)	0.617 (0.605, 0.629)	0.419 (0.339, 0.498)
BJ-ELMLR (Adaptativo)	0.618 (0.605, 0.630)	0.418 (0.341, 0.495)

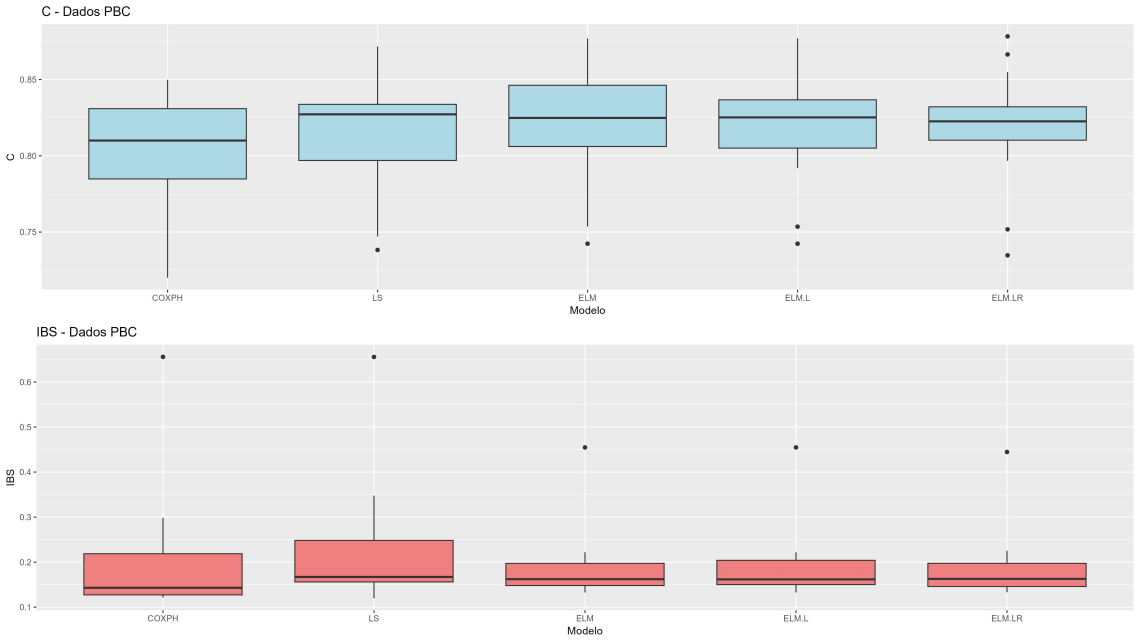

Figura 10 – Pbc: Box-Plot para C-Index e IBS para modelos avaliados.

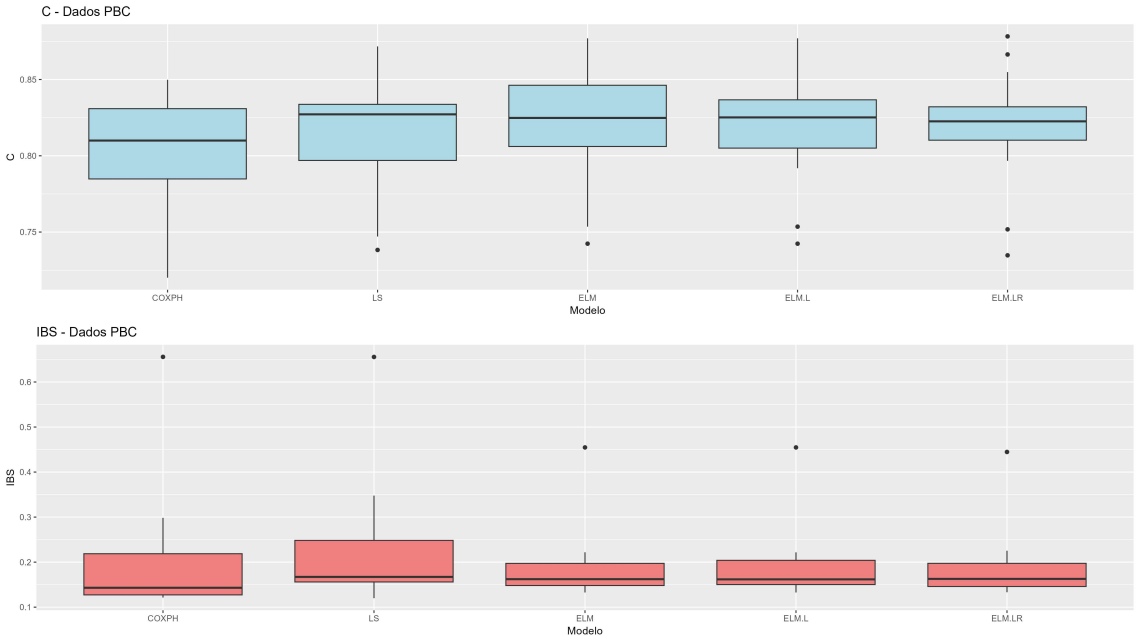

Figura 11 – Lung: Box-Plot para C-Index e IBS para modelos avaliados.

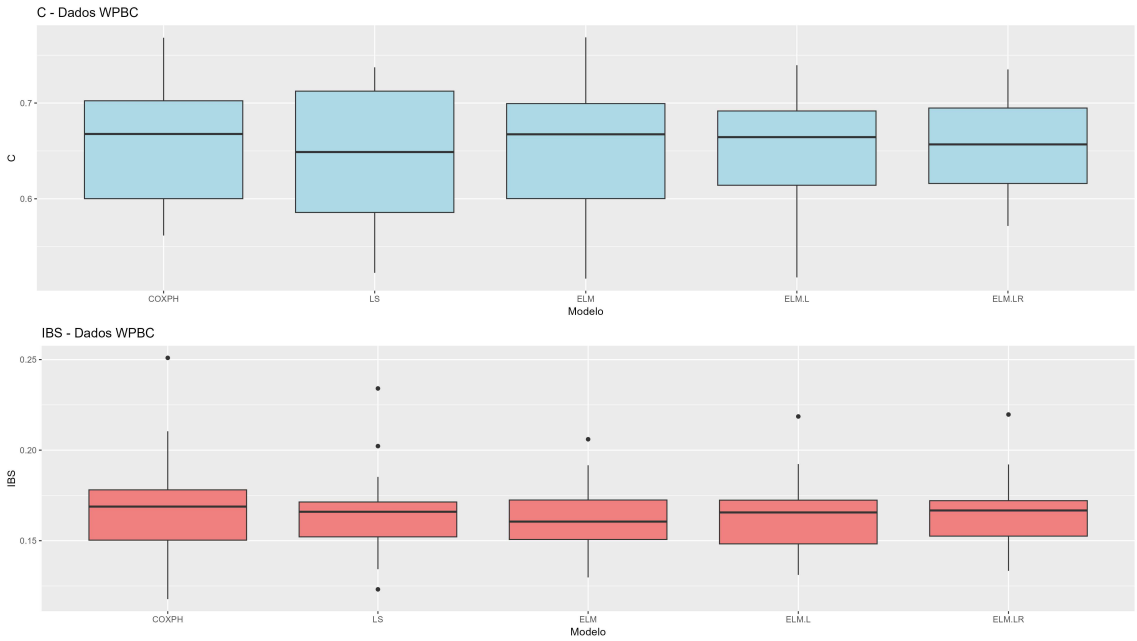

Figura 12 – WPBC: Box-Plot para C-Index e IBS para modelos avaliados.

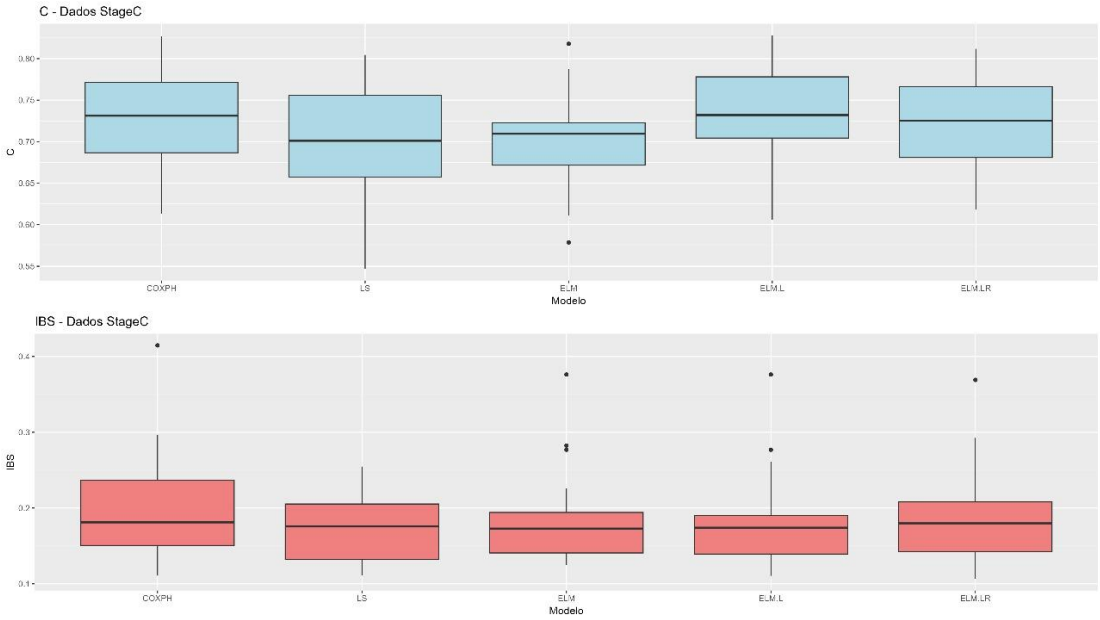

Figura 13 – StageC: Box-Plot para C-Index e IBS para modelos avaliados.

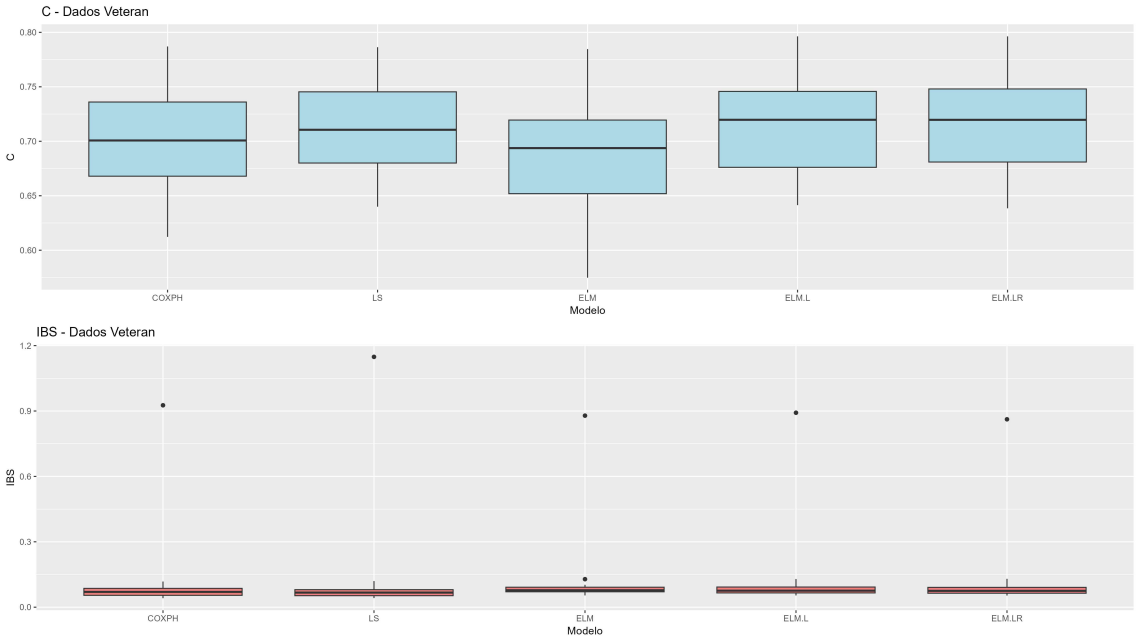

Figura 14 – Veteran: Box-Plot para C-Index e IBS para modelos avaliados.

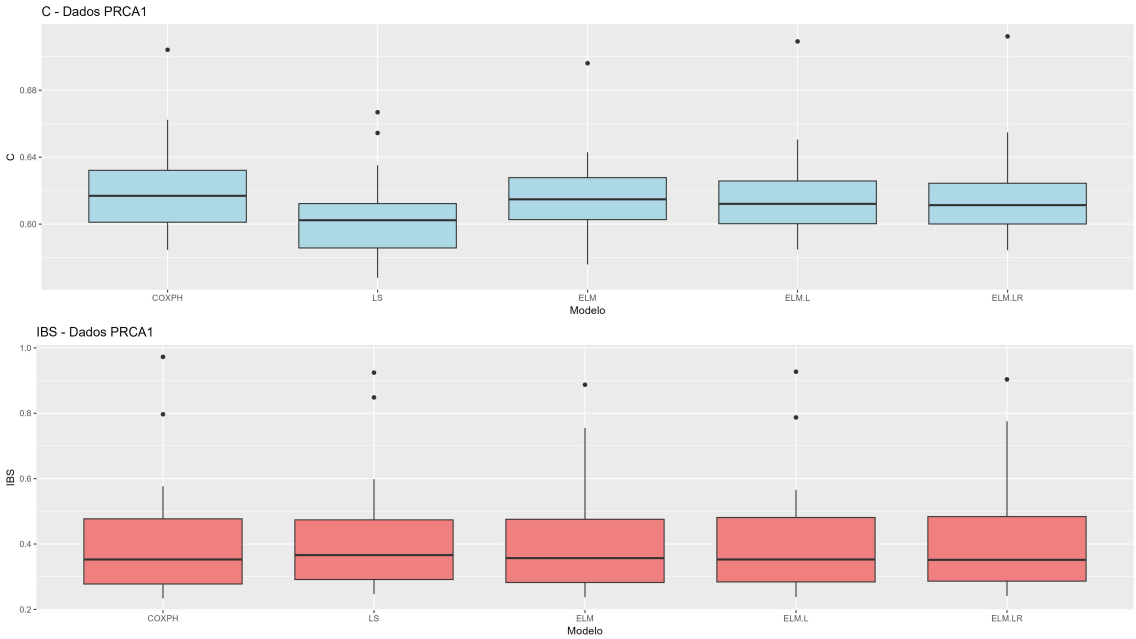

Figura 15 – Prca: Box-Plot para C-Index e IBS para modelos avaliados.

5 DADOS SOBRE DOENÇA RENAL CRÔNICA

Um conjunto de dados referente a Doença Renal Crônica (DRC) foi analisado. A descrição do conjunto de dados e todas as etapas da análise são descritos com detalhe. Primeiramente, descreve-se o estudo que gera o conjunto de dados analisados. Logo apresenta-se o processo de construção do conjunto de dados que será utilizado para as análises desenvolvidas neste trabalho e a avaliação de algumas caraterísticas a partir de diversas análises descritivas. Finalmente, são apresentados os resultados para a análise de sobrevivência que incluem alguns modelos de aprendizado de máquinas para previsão dos tempos de sobrevida são aplicados incluindo o modelo adaptativo proposto neste trabalho. As comparações via C-Index e IBS são apresentados para avaliar os diversos modelos utilizados.

5.1 DESCRIÇÃO DO ESTUDO

O estudo avaliou pacientes com DRC que realizam o tratamento renal de diálise peritonial, desde um estudo delineado como um coorte prospectivo em múltiplos centros. Assim, considerou-se 102 centros distribuídos em todas as regiões geográficas do Brasil que realizam em diálise peritoneal em mais de dez pacientes. Quanto ao tempo de estudo, este foi iniciado em dezembro de 2004 e finalizado em outubro de 2007, desta forma o estudo teve duração de 34 meses contínuos de duração. Durante todo este período foram avaliados 6198 pacientes com tratamento de diálise peritoneal.

O estudo foi submetido ao Comitê de Ética Nacional em Pesquisa Humana e aprovado sob o número 448 e adicionalmente cada clínica ao comitê de ética local. Uma descrição mais detalhada sobre o desenho do estudo, as entidades e pessoas envolvidas, a descrição do processo de obtenção e controle de dados obtidos, descrição dos termos de consentimento, procedimento para registro de informações e outras caraterísticas relevantes do estudo original encontra-se em Suassuna (2009) [51].

O estudo levantou um grande número de variáveis sobre diferentes caraterísticas dos pacientes. A seguir segue descritas estas variáveis:

- Variáveis demográficas: Idade, raça, nível educacional, renda (segundo definição do IBGE), distância até o centro da diálise (Km).
- Variáveis Médicas: Registro de implantação do cateter; complicações relacionadas ao cateter, avaliação do orifício de saída do cateter, volume de ultrafiltrado, teste de ultrafiltração, etiologia de DRC, cuidados pré-dialíticos, história dialítica, comorbidades, medicações em uso, infeções prévias, hospitalização, causa de saída.
- Variáveis sobre qualidade de vida: índice de Karnofsky (medida mensalmente) e questionário SF36 (Medido opcionalmente cada trimestre ou semestre).

- Variáveis de avaliação clínica: Edema, pressão arterial, peso e altura (todas elas medidas mensalmente).
- Variáveis laboratoriais: Ureia, creatinina, ALT, potássio, cálcio, fosfato, glicemia, hemoglobina, hematócrito (todas elas medidas mensalmente), transferrina, ferritina, ferro sérico, albumina, fosfatase alcalina (todas elas medidas trimestralmente), PTHi, anti HBS, HBsAg, anti-HCV, -Kt/v renal e peritoneal (todas elas medidas semestralmente) colesterol total, triglicérides, alumínio sérico, anti-HIV (todas elas medidas anualmente).

A qualidade e a estrutura dos dados foram fundamentais para garantir a robustez das análises subsequentes. O banco de dados em questão apresenta uma heterogeneidade significativa, manifestada através de várias características distintas que são consideradas durante o pré-processamento e a modelagem. Essas características são:

Integridade e Qualidade dos Dados: Por um lado, a presença de dados faltantes (*Missing Data*), observando-se a ocorrência de valores ausentes de forma parcial e aparentemente casual em diversos registros. Por outro lado, a presença de inconsistências e erros de digitação, apresentando registros que contêm inconsistências e valores aberrantes potencialmente resultantes de erros de entrada. A identificação ou remoção destes valores foram essenciais para garantir uma melhor qualidade dos dados.

Natureza das Variáveis: a base de dados incluiu variáveis qualitativas que definem grupos ou categorias, tais como sexo, raça e nível educacional. Estas variáveis exigiram codificação apropriada como a criação de variáveis binárias. Por outro lado, a presença de variáveis contínuas, tais como idade e as diversas variáveis laboratoriais (e.g., pressão arterial, colesterol, triglicerídeos, peso, altura) são quantitativas e contínuas que requerem escalonamento ou normalização para análises comparativas e algoritmos sensíveis à escala.

Unidades de Medida Diversificadas: As variáveis contínuas, particularmente as clínicas e laboratoriais, são expressas em diferentes unidades (e.g., mmHg para pressão arterial, cm para altura, kg para peso, mg/dL para colesterol e triglicerídeos). Esta diversidade impede a comparação direta entre as variáveis e exige padronização ou normalização.

Dinâmica Temporal do Acompanhamento: Datas de início, variável em que os pacientes ingressaram no estudo ao longo do período de coorte definido, resultando em diferentes pontos de partida para o acompanhamento longitudinal de cada indivíduo.

Frequência de Mensuração Heterogênea: A frequência de coleta de dados variou substancialmente entre as variáveis, algumas variáveis foram medidas apenas uma vez, enquanto que outras foram medidas repetidamente em intervalos mensais, trimestrais, semestrais ou anuais.

Data de Término Fixa: O estudo possui uma data final de acompanhamento predeterminada. A combinação do início variável com o término fixo, somada às diferentes

frequências de medição, gera um problema de dados longitudinais desbalanceados. Os pacientes tiveram um número desigual de observações ao longo do tempo, dependendo de quando ingressaram e da frequência de coleta de cada variável.

Censura de Dados de Sobrevida: Para um subconjunto de pacientes acompanhados até outubro de 2007, o tempo exato de sobrevida não é conhecido. Sabe-se apenas que esses indivíduos estavam vivos até aquela data específica (outubro de 2007). Esta situação gera censura à direita.

Considerando as diversas características descritas acima, o tratamento dos dados utilizam diversos mecanismos que dependem do tipo de variável e situação tratada, nos seguintes parágrafos se discutem os mecanismos de uso frequente e que encaixam no tratamento deste banco específico.

Para tratamento de dados faltantes e inconsistências, implementou-se a exclusão de observações incompletas. Paralelamente, inconsistências de digitação foram submetidas a verificação manual, com realização de correções quando aplicáveis.

Na abordagem às variáveis qualitativas, adotou-se codificação *one-hot* para gerar representações binárias por categoria original. Para variáveis quantitativas, devido à diversidade de unidades (mmHg, kg, mg/dL), aplicou-se padronização (z-score) para garantir comparabilidade.

Quanto à complexidade dos dados longitudinais - considerando início variável por paciente, frequências de medição heterogêneas e término fixo do estudo - desenvolveu-se um resumo por paciente através de três métricas sintéticas: primeiro, a média das medições (tendência central); segundo, o desvio padrão (variabilidade intra-paciente); e terceiro, a evolução das medições ao longo do tempo (tendência temporal).

Esta estratégia permite integração eficiente entre trajetórias individuais sintetizadas e variáveis *baseline* medidas uma única vez.

Suassuna (2009) [51] apresenta uma breve caracterização demográfica dos 6.198 pacientes avaliados entre dezembro de 2004 e outubro de 2007. Do total, 277 tinham menos de 12 anos, enquanto 5.921 ultrapassavam essa faixa etária, com média de idade de 58,5 anos (variando de 13 a 101 anos). A distribuição etária revela que 25% dos participantes tinham menos de 48 anos, 50% menos de 60 anos e 75% abaixo de 71 anos, sendo 2.156 idosos acima de 65 anos.

Quanto ao perfil de inclusão no estudo, 2.419 pacientes eram incidentes no tempo zero do registro, e outros 2.281 ingressaram com mais de 90 dias de acompanhamento após o início do programa.

A composição sociodemográfica mostra equilíbrio de gênero: 49,3% eram mulheres. A distribuição étnica aponta predominância de brancos (61%). No aspecto educacional, observou-se que 11,9% eram analfabetos, 52,8% cursaram apenas o ensino fundamental, 22,6% o ensino médio e 7,6% possuíam ensino superior.

Sobre renda familiar mensal, 31,5% recebiam até 2 salários mínimos, 41,6% entre 2 e 5 salários, e 14,8% de 5 a 10 salários. As faixas mais altas mostraram-se menos expressivas: 4,1% declararam renda entre 10 e 20 salários, e apenas 1,2% ultrapassavam 20 salários mínimos.

As análises propostas neste trabalho serão realizadas sobre um conjunto de dados que contém informações que vão além de aspectos demográficos. Incluem-se variáveis relevantes associadas a condições médicas e à qualidade de vida. Dessa forma, torna-se necessário um pré-processamento que considere as características dessas variáveis de maneira conjunta, conforme será descrito nas seções seguintes.

5.2 DEFINIÇÕES PARA PRE-PROCESSAMENTO DE DADOS

Como mencionado por Curioso et al. (2023) [52], a imputação de dados se apresenta como uma abordagem metodologicamente aceita para lidar com valores em falta, sendo uma alternativa ao método de eliminação. No entanto, no contexto específico deste conjunto de dados, optou-se pela aplicação criteriosa da eliminação (*deletion*). Esta decisão foi fundamentalmente sustentada pela dimensão considerável da coorte inicial, que compreendeu 6.198 pacientes estudados. O processo de eliminação adotado, que incluiu a exclusão de registos com valores em falta críticos para a análise, ainda mantém uma base de dados com uma quantidade grande de paciente, mantendo um poder estatístico amplamente suficiente para garantir a validade e robustez das análises realizadas.

A opção pela eliminação é ainda justificada pela natureza do desenho do estudo e pelos objetivos analíticos propostos. Como a análise longitudinal planejada requer acompanhamento temporais para calcular trajetórias individuais dos pacientes a partir dos três indicadores propostos, considera-se que imputação de valores faltantes introduziria distorções indesejadas nos dados.

Para caracterizar os dados longitudinais de cada paciente, foram definidos três indicadores principais:

1. **Tendência central** — representada pela média ou, de forma mais robusta, pela mediana dos registros. A mediana é preferida em séries clínicas devido à sua resistência a valores extremos e variações abruptas.

2. **Dispersão interna** — medida pelo coeficiente de variação (desvio padrão/média). Como alternativa robusta, utilizou-se a razão entre o intervalo interquartílico (Q3–Q1) e a mediana, que reduz a influência de *outliers* e distribuições assimétricas.

3. **Tendência temporal** — estimada pela inclinação de um modelo de regressão linear simples ou, de forma mais robusta, pela mediana das diferenças sucessivas ($X_{i+1} - X_i$). Esta última abordagem é especialmente adequada para séries curtas

(entre 5 e 15 observações por paciente) e para dados clínicos sujeitos a *outliers* ou variações técnicas.

A escolha por indicadores robustos é sustentada pela literatura: Wilcox [53] destaca que estimadores baseados em medianas são menos sensíveis a observações atípicas, enquanto Tukey [54] recomenda o uso de diferenças sucessivas como ferramenta exploratória em séries temporais. Dessa forma, o conjunto de indicadores adotado permite capturar valores típicos, variabilidade relativa e tendências clínicas relevantes sem comprometer a estabilidade estatística das estimativas.

5.3 CONSTRUÇÃO DO ARQUIVO PARA ANÁLISE

Os dados originais da pesquisa encontram-se em arquivos separados por tipos de informações, desta forma eles foram tratados separadamente. O arquivo final para análise foi construído a partir da busca de informações nos arquivos originais do estudo e ligadas pelo registro de ID do paciente presente em cada um dos arquivos originais, as variáveis selecionadas para análise foi sugerida por um especialista em nefrologia, que devam incluir os fatores de risco clássicos para DRC, tais como idade, comorbidades como diabetes e doenças cardiovasculares, e resultados de medições de exames médicos e laboratoriais.

O processo de integração de dados iniciou-se com o processamento de dados imutáveis, padronizando variáveis alfanuméricas e atribuindo uma codificação numérica quando necessário. O resultado final deste processo é o registro de 6.128 pacientes com variáveis demográficas essenciais: Idade, Sexo, Raça, Instrução, Renda, entre outras. Esse procedimento originou um primeiro arquivo composto por variáveis classificadas como imutáveis, por terem sido coletadas apenas uma vez ao longo do estudo.

Para o processamento das informações de óbito, filtrou-se o último registro de cada paciente para identificar o desfecho (óbito ou outras formas de saídas). Após a avaliação de informações, obteve-se registros para 5.870 pacientes com as variável Motivo de Saída, utilizada para definir a informação de censura. A partir desse processamento, foi gerado um arquivo que inclui tanto o indicador de censura quanto o tempo de sobrevida correspondente.

Os registros de comorbidades associadas à DRC encontram-se em outro arquivo, neste caso foi feita a conversão das respostas qualitativas para valores numéricos. Após consolidar comorbidades individuais em um único arquivo, obteve-se um registro de 5.603 pacientes e variáveis ID do paciente, Diabetes, Doença Cardiovascular (DCV) e Hipertensão Arterial Sistêmica (HAS).

A avaliação de qualidade de vida é obtida a partir de cálculo de escores como Capacidade Funcional (CF), Saúde Mental (MH) desde o questionário denominado SF36 e o índice de Karnofsky. Logo, eles foram validados comparando com as datas de avaliação

da enfermagem eliminado registros sem data de registro. Finalmente, agregaram-se médias/medianas por paciente, criando um arquivo de 5.892 pacientes com métricas sumarizadas de Karnofsky e domínios do questionário SF36.

Os registros de dados clínicos (tais como pressão arterial sistólica ou diastólica) e laboratoriais (como fósforo e potássio) apresentam medições longitudinais. O processamento consistiu na conversão de valores alfanuméricos em códigos numéricos. Considerando a quantidade limitada de observações longitudinais por paciente e a possibilidade de registros extremos decorrentes de erros clínicos ou técnicos, optou-se pelo uso exclusivo de indicadores robustos, sintetizados da seguinte forma:

- Tendência central — representada pela mediana dos registros, em substituição à média, por sua maior resistência a valores atípicos;
- Dispersão interna — expressa pela razão entre o intervalo interquartílico (Q3–Q1) e a mediana, alternativa robusta ao coeficiente de variação tradicional;
- Tendência temporal — definida pela mediana das diferenças entre medições sucessivas ($X_{i+1} - X_i$), em vez da regressão linear, dada a curta extensão das séries e o caráter ordinal dos registros temporais.

Desta forma, buscou-se não apenas identificar o valor típico e a variabilidade intrínseca, mas também quantificar a direção e a taxa de mudança de cada parâmetro clínico ao longo do acompanhamento. Ao final desse processo, foi gerado um arquivo contendo 5.697 pacientes com os indicadores sugeridos.

Finalmente, o arquivo final para análise unifica os dados provenientes de todos os processos descritos, integrados por meio da variável de identificação do paciente (ID do paciente). Dessa forma, o arquivo consolidado reúne informação sobre tempo de sobrevida, censura, desfechos, dados demográficos, comorbidades, métricas de qualidade de vida e exames clínicos e laboratoriais.

A Figura 16 apresenta o fluxograma que ilustra as etapas descritas indicando a quantidade de dados excluídos.

Observa-se que o resultado da fusão destes dados inclui registros de 6.198 pacientes. No entanto, foi necessária a aplicação de filtros de exclusão consecutivos: (1) 1.381 pacientes sem exames clínicos e laboratoriais; (2) 145 pacientes sem registro de sexo; (3) 61 pacientes sem registro de idade; e (4) 251 com idade menor que 18 anos. Assim, o arquivo final contém dados de 4.360 pacientes disponíveis para as análises, o que corresponde ao 70,35% do total de participantes do estudo.

É importante ressaltar que as exclusões foram baseadas na ausência de dados em variáveis consideradas relevantes, a principal causa de exclusão foi a ausência de informação de exames clínicos e laboratoriais, que corresponde a 22,28% do total (75,51%

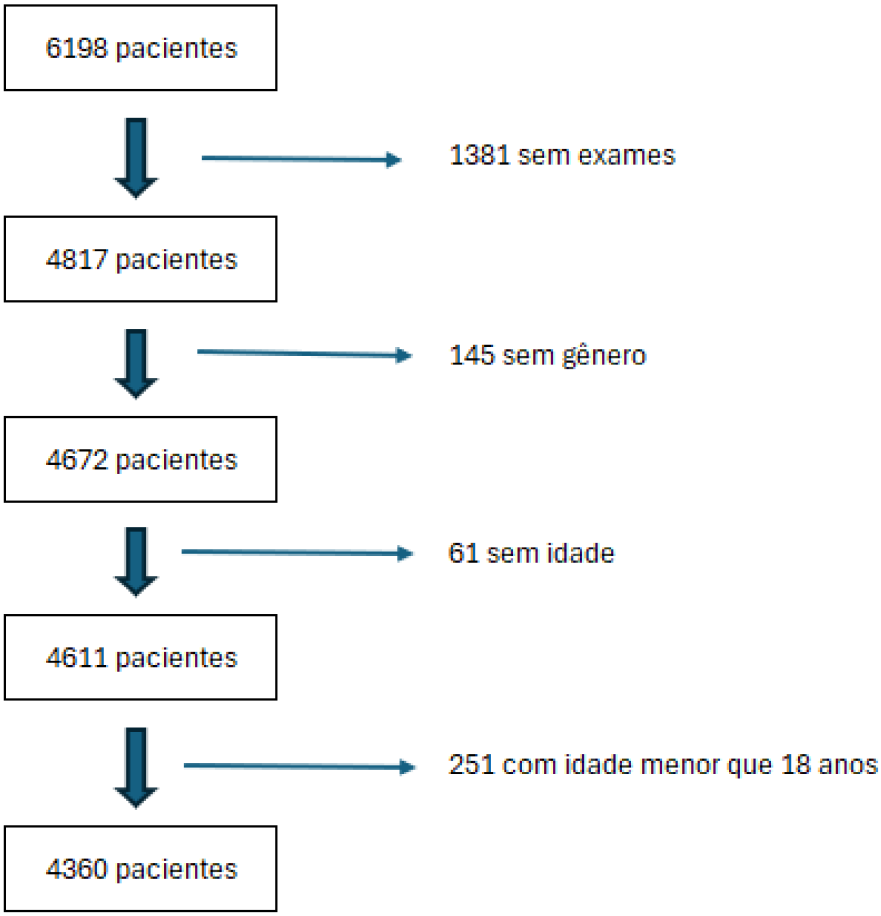

Figura 16 – Fluxograma para construção de banco de análise.

das eliminações). As demais exclusões decorreram da ausência de registro de sexo ou idade. Além disso, optou-se por não incluir pacientes menores de idade.

5.4 DESCRIÇÃO DE DADOS PARA ANÁLISE

Nesta seção apresenta-se uma descrição das varáveis consideradas para a análise de sobrevivência, provenientes do banco de pacientes com DRC em tratamento por diálise peritonial. Foi considerado o tempo de sobrevida atê o desfecho, que neste corresponde ao óbito. As covariáveis incluídas são os fatores clássicos associados a este tipo de doença. A seguir descrevem-se as variáveis analisadas:

No contexto da modelagem de sobrevivência, duas variáveis centrais estruturam a análise:

- **Tempo de sobrevida**, Neste caso, a variável corresponde ao tempo de sobrevida do paciente com DRC após o início do tratamento por diálise peritoneal. Considerando que a duração do estudo é de 34 meses e que, segundo o protocolo, são admitidos

pacientes com até três meses em diálise peritoneal, os valores admissíveis para essa variável variam de 1 a 37 meses.

A Figura 17 apresenta os histogramas das distribuições relativas para o tempo de sobrevida e para sua transformação logarítmica. Observa-se que a distribuição do tempo de sobrevida apresenta assimetria à direita, característica comum em variáveis relacionadas à duração de eventos clínicos. Essa assimetria é evidenciada pela discrepância entre a forma empírica da distribuição e a curva normal teórica sobreposta no gráfico.

A transformação logarítmica, exigida pelo modelo Buckley-James, tem como objetivo aproximar a distribuição dos dados de uma forma mais simétrica. Essa aproximação é visível no histograma da direita, onde a distribuição transformada se apresenta mais próxima da curva normal teórica. No entanto, mesmo após a transformação, persistem discrepâncias nos extremos da distribuição em relação ao comportamento esperado sob normalidade. A adoção de modelos robustos visa mitigar essas diferenças, oferecendo maior flexibilidade para acomodar desvios nos dados observados.

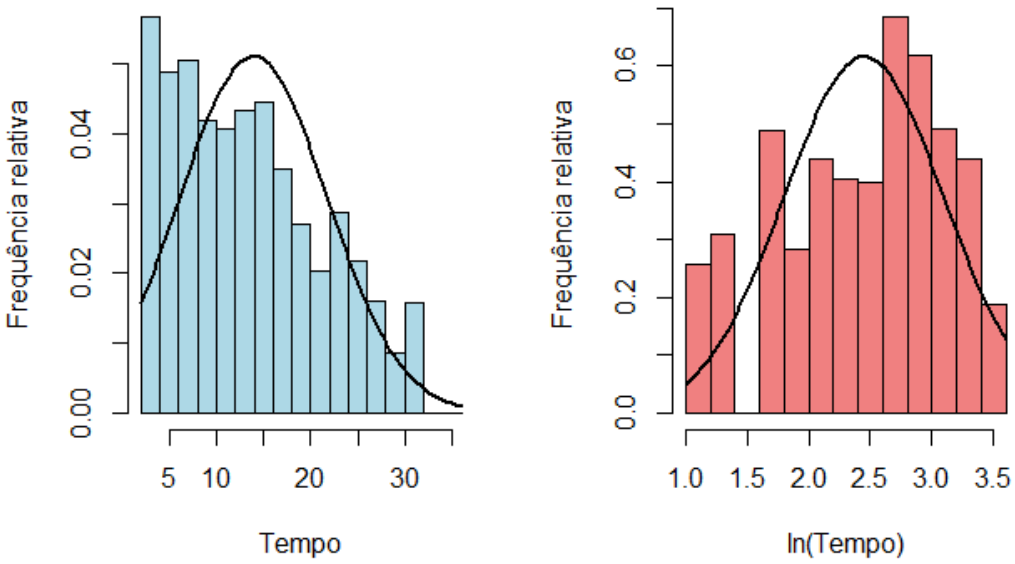

Figura 17 – Histogramas comparativos: distribuição dos tempos de sobrevida e da sua transformação logarítmica.

- **Motivo de saída**, variável que define a ocorrência de eventos ao longo do estudo e a partir da qual é estabelecido o registro de censura. Esta variável foi categorizada em três situações distintas:

  1. **Óbito**, representando o desfecho principal de interesse.

2. **Saída antecipada**, caracterizada pela interrupção do acompanhamento sem ocorrência de óbito. Essa situação pode decorrer de diferentes motivos, como recuperação da função renal, transferência para hemodiálise ou realização de transplante renal.
3. **Conclusão do estudo vivo**, indicando indivíduos que permaneceram em acompanhamento até o final do período estabelecido sem ocorrência de óbito.

Com base na variável anterior, foi definida a variável **status** para o modelo de sobrevivência, cujo desfecho de interesse é o óbito. Trata-se de uma variável binária que indica se o evento de interesse foi observado ou não. Indivíduos que apresentaram óbito durante o período de acompanhamento são classificados como não censurados (valor 1), enquanto aqueles que encerraram o acompanhamento por saída antecipada ou por conclusão do estudo vivos são classificados como censurados (valor 0).

A definição de censura considera que a saída antecipada corresponde a uma interrupção do acompanhamento sem ocorrência do evento de interesse, podendo decorrer de causas clínicas ou administrativas, como recuperação da função renal, transferência para hemodiálise ou realização de transplante. Já a conclusão do estudo vivo refere-se aos participantes que permaneceram em acompanhamento até o final do período estabelecido sem registro de óbito. Em ambos os casos, o tempo de sobrevida observado é incompleto em relação ao tempo total até o evento, justificando a classificação como censura. No presente estudo, 78,3% dos dados foram censurados, sendo que a maioria destes (62,9%) correspondeu a participantes que concluíram o estudo vivos. A Tabela 22 apresenta estas quantificações.

Tabela 22 – Distribuição da variável Status e motivo de saída do estudo

Status	**Motivo de saída**	**n**	**%**
Óbito	—	947	21,7%
Censura	Saída antecipada	669	15,3%
	Conclusão do estudo vivo	2744	62,9%
	Total censura	**3413**	**78,3%**
Total geral	—	**4360**	**100,0%**

A Tabela 23 apresenta os valores de tempo de sobrevida observados no estudo. Como discutido anteriormente, é possível encontrar registros de sobrevida até 37 meses, embora o acompanhamento tenha duração máxima de 34 meses. Observa-se que os pacientes que evoluíram para óbito e aqueles com saída antecipada apresentam comportamento semelhante, enquanto os indivíduos que concluíram o estudo vivos exibem tempos mais elevados, como esperado. Verifica-se assimetria à direita em todas as categorias, caracterís-

tica recorrente em estudos clínicos de sobrevida, refletindo a presença de poucos pacientes que sobrevivem por períodos mais longos.

Tabela 23 – Medidas resumo para Tempo de sobrevida segundo motivo de saída

Estatística	**Óbito**	**Saída antecipada**	**Conclusão do estudo vivo**	**Total**
Média	11,03	11,04	15,61	13,92
Mediana	10,00	9,00	15,00	13,00
Desvio padrão	6,68	6,50	7,99	7,82
Mínimo	1,00	3,00	3,00	1,00
Percentil 25	6,00	6,00	9,00	7,00
Percentil 75	15,00	15,00	22,00	19,00
Máximo	36,00	31,00	32,00	36,00

As covariáveis incluídas correspondem às mais frequentemente analisadas em estudos clínicos de DRC (veja, por exemplo, Suassuna (2009) [51]) e são reconhecidas como relevantes por sua associação com a progressão da doença. O conjunto considerado foi validado por profissionais dedicados ao manejo da DRC, em especial no contexto da diálise peritoneal.

A análise descritiva que se segue permitirá caracterizar o perfil da população estudada, fornecendo uma visão abrangente das distribuições das variáveis e estabelecendo a base para a modelagem estatística subsequente. A apresentação das frequências e medidas de tendência central e dispersão possibilitará identificar padrões e heterogeneidades iniciais entre os fatores de risco e os desfechos de interesse.

5.4.1 Características sociodemográficas

As variáveis sociodemográficas consideradas para a análise incluem **raça/cor**, **renda familiar** e **idade**. Elas descrevem o perfil básico da população estudada e permitem compreender a composição da coorte. As variáveis raça/cor e renda familiar são qualitativas.

A distribuição por **raça/cor** mostra predominância de indivíduos brancos (63,9%), seguidos por pardos (22,0%) e pretos (11,3%). As categorias amarela (2,9%) e indígena (0,1%) apresentam baixa representatividade, motivo pelo qual podem ser agrupadas em uma categoria denominada “outros”.

Em relação à **renda familiar**, observa-se maior concentração na faixa de 2 a 5 salários mínimos (44,2%), seguida por até 2 salários mínimos (32,0%). Faixas mais altas de rendimento (acima de 10 salários mínimos) são pouco frequentes, representando cerca de 6,5% do total. Por essa razão, foi criada uma categoria conjugada: “acima de 5 salários mínimos”. Esses resultados indicam que a coorte é composta majoritariamente

por indivíduos de renda baixa a média e de cor branca, com participação relevante de pardos e pretos.

Os resultados, segundo motivo de saída, são apresentados na Tabela 24. Em todas as situações de saída, as distribuições de frequência se mostram similares. Para a análise de sobrevivência, cada categoria será considerada uma variável dicotômica. Em cada variável, será desconsiderada a categoria com menor frequência (sem rendimento para renda e "outros" para raça/cor), a fim de evitar problemas teóricos de colinearidade e singularidade da matriz.

Tabela 24 – Distribuição percentual para variáveis renda e raça/cor segundo motivo de saída

	Óbito	**Saída antecipada**	**Conclusão do estudo vivo**	**Total**
Renda				
Sem rendimento	1,0%	1,0%	1,3%	1,3%
Até 2	36,0%	32,7%	30,4%	32,0%
2 a 5	39,6%	42,2%	46,4%	44,2%
5 a 10	16,9%	17,5%	16,3%	16,5%
10 a 20	5,0%	5,6%	5,4%	5,3%
Maior 20	1,6%	0,6%	1,2%	1,2%
Raça/cor				
Branco	66,6%	68,6%	62,2%	63,9%
Preto	11,6%	9,4%	11,6%	11,3%
Pardo	18,9%	22,1%	22,8%	22,0%
Amarelo	2,7%	0,7%	3,3%	2,9%
Indígena	0,1%	0,1%	0,1%	0,1%
TOTAL	**100%**	**100%**	**100%**	**100%**

Tabela 25 – Medidas resumo para Idade e segundo motivo de saída

Estatística	**Óbito**	**Saída antecipada**	**Conclusão do estudo vivo**	**Total**
Média	66,90	54,60	57,58	59,15
Mediana	68,00	55,00	59,00	60,00
Desvio padrão	14,13	15,95	15,81	16,04
Mínimo	20,00	18,00	18,00	18,00
Percentil 25	58,00	43,00	47,00	49,00
Percentil 75	77,00	67,00	69,00	71,00
Máximo	97,00	95,00	101,00	101,00

A Tabela 25 apresenta as medidas da variável idade segundo o motivo de saída, evidenciando diferenças marcantes entre os grupos. Pacientes que evoluíram para óbito apresentam média de 66,9 anos, superior àqueles que tiveram saída antecipada (54,6 anos) ou concluíram o estudo vivos (57,6 anos).

A mediana acompanha esse padrão, reforçando que o risco de óbito está associado a maior idade. O intervalo observado é amplo (18 a 101 anos), mas os percentis mostram que três quartos dos pacientes que faleceram tinham mais de 58 anos, enquanto nos demais grupos predominam idades mais jovens. Esses achados confirmam a relevância da idade como fator de risco e justificam sua inclusão como covariável na análise de sobrevivência.

5.4.2 Escala de Desempenho de Karnofsky

A **Escala de Desempenho de Karnofsky** [55] é uma medida clínica desenvolvida originalmente em 1949 para avaliar o estado funcional de pacientes, especialmente em oncologia, mas também aplicada em outras áreas como nefrologia e cuidados paliativos. A escala varia de 0 a 100, em que valores mais altos indicam maior independência e capacidade de realizar atividades diárias, enquanto valores mais baixos refletem dependência significativa ou óbito. Trata-se de um instrumento amplamente utilizado para estimar prognóstico e orientar decisões terapêuticas.

A aplicação da escala é realizada por médicos ou profissionais de saúde treinados, que avaliam a condição geral do paciente considerando sintomas, necessidade de assistência e capacidade de autocuidado. Por exemplo, um paciente com escore 80 ainda consegue realizar atividades normais, embora com esforço e sintomas leves, enquanto um paciente com escore 40 já apresenta incapacidade funcional importante e requer cuidados contínuos. Essa avaliação, embora subjetiva, é padronizada e validada, o que garante sua utilidade clínica.

O uso da Escala de Karnofsky permite comparar a efetividade de diferentes terapias, estimar a tolerância a tratamentos como quimioterapia, radioterapia ou diálise e identificar pacientes que necessitam de suporte intensivo. Por sua simplicidade e relevância prognóstica, essa escala permanece como uma ferramenta clássica e essencial em estudos clínicos e na prática médica. Como comentário adicional, o estudo também levantou outras avaliações de qualidade de vida; no entanto, elas foram descartadas como covariáveis, pois 1.245 pacientes (28,6%) não apresentavam essas informações.

A Tabela 26 evidencia diferenças relevantes na Escala de Karnofsky segundo o motivo de saída. Pacientes que evoluíram para óbito apresentam média de 70,1 pontos, inferior àqueles que tiveram saída antecipada (81,7) ou concluíram o estudo vivos (81,4), indicando pior desempenho funcional associado ao desfecho mais grave. A mediana e os percentis confirmam esse padrão, mostrando que três quartos dos pacientes que faleceram tinham escore abaixo de 80, enquanto nos demais grupos predominam valores mais elevados.

O desvio padrão também sugere maior heterogeneidade entre os pacientes que foram a óbito. Esses resultados reforçam a associação entre menor desempenho funcional e maior risco de mortalidade, justificando a inclusão da Escala de Karnofsky como covariável na análise de sobrevivência.

Tabela 26 – Medidas resumo para Escala de Karnofsky segundo motivo de saída

Estatística	**Óbito**	**Saída antecipada**	**Conclusão do estudo vivo**	**Total**
Média	70,11	81,65	81,40	78,99
Mediana	71,25	83,33	82,86	80,83
Desvio padrão	14,41	12,16	11,40	13,00
Mínimo	12,50	27,00	25,00	12,50
Percentil 25	60,40	75,00	75,00	71,11
Percentil 75	80,00	90,00	90,00	88,00
Máximo	100,00	100,00	100,00	100,00

5.4.3 Comorbidades

A presença de comorbidades é um aspecto central na análise de sobrevivência em pacientes com DCR, pois condições associadas como doença cardiovascular (DCV), diabetes mellitus e hipertensão arterial sistêmica (HAS) são reconhecidas como fatores de risco bem estabelecidos para progressão da doença e aumento da mortalidade. Essas variáveis permitem caracterizar o perfil clínico da população e avaliar sua influência sobre os desfechos, justificando sua inclusão como covariáveis no modelo estatístico.

Tabela 27 – Prevalência de comorbidades segundo motivo de saída

	Óbito	**Saída antecipada**	**Conclusão do estudo vivo**	**Total**
DCV	70,7%	61,9%	60,2%	62,7%
Diabetes	45,8%	34,8%	39,1%	39,9%
HAS	16,1%	16,6%	16,8%	16,6%

A Tabela 27 apresenta a prevalência das comorbidades segundo o motivo de saída. Observa-se que a DCV é mais frequente entre os pacientes que evoluíram para óbito (70,7%) em comparação com os demais grupos (cerca de 61%). O diabetes também mostra maior prevalência entre os óbitos (45,8%) em relação aos pacientes que permaneceram no estudo ou não faleceram (aproximadamente 35 e 39%). Já a HAS apresenta distribuição relativamente homogênea entre os grupos, em torno de 16%, sem diferenças marcantes. Esses resultados reforçam a associação entre DCV e diabetes com maior risco de mortalidade, enquanto a

HAS, embora relevante como fator de risco, não se diferencia de forma expressiva entre os desfechos.

5.4.4 Variáveis clínicas e laboratoriais

Além das comorbidades, foram incluídas no modelo variáveis clínicas e laboratoriais com potencial impacto sobre a evolução da doença renal crônica (DRC). A variável clínica considerada foi a **pressão arterial sistólica (PAS)**, selecionada em detrimento da diastólica devido à sua maior relevância epidemiológica e clínica. Estudos apontam a PAS como um marcador mais sensível de risco cardiovascular e renal, especialmente em populações com comprometimento da função renal, sendo fortemente associada à progressão da DRC e à mortalidade.

As variáveis laboratoriais selecionadas foram **hemoglobina (HEM)**, **potássio (POT)** e **fósforo (FOS)**, todas com implicações diretas no estado metabólico e na estabilidade clínica dos pacientes. A hemoglobina é um indicador da presença e gravidade da anemia, condição comum em pacientes renais e associada a pior prognóstico. O potássio, por sua vez, reflete o equilíbrio eletrolítico e está relacionado a complicações cardiovasculares graves, como arritmias, especialmente em casos de hiperpotassemia. Já o fósforo é um marcador importante do metabolismo mineral, cuja elevação está associada à calcificação vascular, disfunção endotelial e aumento do risco cardiovascular em pacientes com DRC.

Essas variáveis foram monitoradas longitudinalmente ao longo do estudo. Para lidar com a limitação de registros por paciente e a presença de valores extremos, optou-se pelo uso de medidas robustas capazes de sintetizar os dados de forma confiável. Conforme justificado na seção anterior, foram aplicadas propostas robustas para tendência central, dispersão interna e tendência temporal.

A verificação de dados faltantes e de registros com valores fora dos limites plausíveis resultou na exclusão de 18 casos, o que corresponde a 0,4% do total de 4.360 pacientes. Dessa forma, as análises subsequentes foram realizadas com 4.342 observações válidas.

Os resultados apresentados na Tabela 28 evidenciam diferenças relevantes na pressão arterial sistólica entre os grupos de saída.

Na dimensão de tendência central, observa-se que os pacientes que evoluíram para óbito apresentaram valores médios e medianos mais baixos (130 mmHg), enquanto aqueles que permaneceram no estudo ou tiveram saída antecipada registraram valores mais elevados (em torno de 137–140 mmHg). Esse resultado sugere que níveis mais baixos de PAS podem estar associados a maior risco de mortalidade, embora também reflitam a heterogeneidade clínica da população.

Quanto à dispersão interna, os pacientes que faleceram apresentaram maior variabilidade relativa (média de 19,06), em comparação com os demais grupos (cerca de 17).

Tabela 28 – Medidas resumo para a variável Pressão Arterial Sistólica (PAS) segundo motivo de saída

Estatística	Óbito	Saída antecipada	Conclusão do estudo vivo	Total
Tendência central				
Média	130,63	137,69	137,22	135,87
Mediana	130,00	140,00	140,00	139,25
Desvio padrão	21,77	19,23	18,46	19,53
Mínimo	80,00	78,00	78,00	78,00
Percentil 25	120,00	125,00	125,00	120,00
Percentil 75	140,00	150,00	150,00	150,00
Máximo	230,00	240,00	225,00	240,00
Dispersão interna				
Média	19,06	17,10	17,07	17,90
Mediana	16,67	15,38	15,86	16,41
Desvio padrão	13,30	10,00	10,58	10,69
Mínimo	0,00	0,00	0,00	0,00
Percentil 25	9,09	8,33	11,11	10,53
Percentil 75	25,00	23,08	23,08	23,35
Máximo	125,00	83,33	80,00	125,00
Tendência temporal				
Média	-1,51	-1,22	-0,24	-0,67
Mediana	0,00	0,00	0,00	0,00
Desvio padrão	8,05	7,07	5,82	6,58
Mínimo	-50,00	-45,00	-40,00	-50,00
Percentil 25	-2,00	0,00	0,00	0,00
Percentil 75	0,00	0,00	0,00	0,00
Máximo	52,00	30,00	35,00	52,00

Essa maior amplitude pode indicar instabilidade pressórica, frequentemente relacionada a pior prognóstico em indivíduos com DRC.

Na dimensão de tendência temporal, a mediana das diferenças sucessivas foi nula em todos os grupos, indicando ausência de tendência sistemática de aumento ou redução da PAS ao longo do tempo. Entretanto, os valores extremos revelam oscilações importantes, especialmente entre os pacientes que foram a óbito, com variações negativas acentuadas (mínimo de –50 mmHg). Essa instabilidade temporal reforça o papel da pressão arterial como marcador dinâmico de risco, cuja flutuação pode impactar diretamente a sobrevivência.

Em conjunto, os resultados mostram que a PAS não apenas difere em níveis médios entre os desfechos, mas também apresenta padrões de variabilidade e instabilidade temporal que devem ser considerados na modelagem estatística e na interpretação clínica da progressão da DRC.

Tabela 29 – Medidas resumo para a variável Hemoglobina (HEM) segundo motivo de saída

Estatística	**Óbito**	**Saída antecipada**	**Conclusão do estudo vivo**	**Total**
Tendência central				
Média	11,26	11,26	11,42	11,36
Mediana	11,30	11,30	11,50	11,40
Desvio padrão	1,66	1,65	1,58	1,61
Mínimo	5,70	6,20	6,00	5,05
Percentil 25	10,20	10,18	10,45	10,30
Percentil 75	12,30	12,30	12,40	12,30
Máximo	16,90	18,30	23,05	23,05
Dispersão interna				
Média	16,26	17,09	15,81	16,11
Mediana	13,48	13,98	13,24	13,70
Desvio padrão	12,64	15,35	10,54	11,87
Mínimo	0,00	0,00	0,00	0,00
Percentil 25	8,89	8,41	9,08	8,94
Percentil 75	19,91	21,80	19,92	20,00
Máximo	152,54	180,71	164,76	180,71
Tendência temporal				
Média	0,05	0,05	0,05	0,05
Mediana	0,00	0,00	0,00	0,00
Desvio padrão	0,57	0,93	0,48	0,59
Mínimo	-5,70	-3,20	-8,65	-8,65
Percentil 25	-0,10	-0,20	-0,10	-0,10
Percentil 75	0,20	1,20	0,40	0,40
Máximo	2,90	12,95	4,20	12,95

A Tabela 29 evidencia o comportamento da hemoglobina segundo o motivo de saída dos pacientes.

Na dimensão de tendência central, os valores médios e medianos são bastante próximos entre os grupos, variando em torno de 11,3 a 11,4 g/dL. Isso indica que, em termos gerais, os níveis de hemoglobina se mantêm relativamente estáveis independentemente do desfecho. Contudo, os valores mínimos observados (entre 5 e 6 g/dL) revelam a presença

de casos de anemia significativa, condição sabidamente associada à pior evolução clínica em pacientes com DCR.

Quanto à dispersão interna, nota-se maior variabilidade nos grupos de saída antecipada e óbito, com desvios padrão mais elevados e amplitudes maiores. Essa maior heterogeneidade sugere que oscilações nos níveis de hemoglobina podem estar relacionadas a maior instabilidade clínica e risco de complicações. Já o grupo que concluiu o estudo vivo apresenta menor dispersão, o que pode refletir maior estabilidade hematológica.

Na dimensão de tendência temporal, a mediana das diferenças sucessivas é nula em todos os grupos, indicando ausência de tendência sistemática de aumento ou redução da hemoglobina ao longo do tempo. Entretanto, os valores extremos revelam quedas acentuadas em alguns pacientes (mínimos de até –8,65 g/dL), especialmente entre os que faleceram, o que reforça a importância da monitorização contínua da anemia.

Em síntese, os resultados mostram que, embora os níveis médios de hemoglobina sejam semelhantes entre os grupos, a variabilidade interna e as oscilações temporais desempenham papel relevante na caracterização do risco, justificando a inclusão dessa variável como covariável na análise de sobrevivência.

A Tabela 30 apresenta os resultados descritivos para o potássio, segundo o motivo de saída dos pacientes.

Na dimensão de tendência central, os valores médios e medianos se mantêm próximos entre os grupos, variando em torno de 4,1 a 4,3 mEq/L, dentro da faixa considerada fisiológica. Entretanto, os valores máximos observados (até 7,4 mEq/L) indicam episódios de hiperpotassemia, condição crítica em pacientes com DRC, por estar associada a risco elevado de arritmias e mortalidade súbita.

Quanto à dispersão interna, observa-se grande variabilidade, especialmente nos grupos de saída antecipada e conclusão do estudo vivo, com desvios padrão elevados e máximos extremos (até 935 mEq/L). Esses valores refletem a presença de registros atípicos ou erros de medição, mas também sugerem que oscilações significativas nos níveis de potássio podem ocorrer ao longo do acompanhamento. O grupo de óbito apresenta menor amplitude, mas ainda com variabilidade relevante, reforçando a importância do controle rigoroso desse parâmetro.

Na dimensão de tendência temporal, a mediana das diferenças sucessivas é nula em todos os grupos, indicando ausência de tendência sistemática de aumento ou redução do potássio ao longo do tempo. Contudo, os valores mínimos negativos (até –9,35 mEq/L) e máximos positivos (até +1,45 mEq/L) revelam oscilações pontuais que podem refletir tanto ajustes terapêuticos quanto instabilidade clínica.

Em síntese, os resultados mostram que, embora os níveis médios de potássio se mantenham dentro da faixa fisiológica, a variabilidade interna e as oscilações temporais desempenham papel crítico na caracterização do risco. Isso justifica a inclusão do potássio

Tabela 30 – Medidas resumo para a variável Potássio (POT) segundo motivo de saída

Estatística	Óbito	Saída antecipada	Conclusão do estudo vivo	Total
Tendência central				
Média	4,14	4,37	4,36	4,31
Mediana	4,10	4,25	4,31	4,25
Desvio padrão	0,74	0,71	0,64	0,68
Mínimo	2,10	2,25	2,10	2,10
Percentil 25	3,60	3,60	3,91	3,85
Percentil 75	4,70	4,85	4,85	4,85
Máximo	6,80	7,40	6,80	7,40
Dispersão interna				
Média	19,54	18,25	18,02	18,38
Mediana	16,67	15,56	15,91	15,91
Desvio padrão	13,38	19,87	20,40	20,21
Mínimo	0,00	0,00	0,00	0,00
Percentil 25	10,87	10,14	11,27	11,11
Percentil 75	25,00	22,77	21,66	22,62
Máximo	102,60	444,19	935,29	935,29
Tendência temporal				
Média	-0,04	-0,02	-0,01	-0,02
Mediana	0,00	0,00	0,00	0,00
Desvio padrão	0,26	0,44	0,21	0,27
Mínimo	-1,40	-9,35	-1,70	-9,35
Percentil 25	0,00	0,00	0,00	0,00
Percentil 75	0,05	0,10	0,05	0,05
Máximo	1,45	1,15	1,40	1,45

como covariável na análise de sobrevivência, dada sua relevância clínica para complicações cardiovasculares e desfechos adversos em pacientes com DRC.

A Tabela 31 apresenta os resultados descritivos para o fósforo, segundo o motivo de saída dos pacientes.

Na dimensão de tendência central, observa-se que os valores médios e medianos se mantêm próximos entre os grupos, variando em torno de 4,6 a 5,0 mg/dL, dentro da faixa considerada aceitável para pacientes em acompanhamento. Contudo, os valores máximos (até 13,2 mg/dL) indicam episódios de hiperfosfatemia, condição associada à calcificação vascular, disfunção endotelial e maior risco cardiovascular em indivíduos com DRC.

Quanto à dispersão interna, nota-se variabilidade significativa em todos os grupos,

Tabela 31 – Medidas resumo para variável Fósforo (FOS) segundo motivo de saída

Estatística	Óbito	Saída antecipada	Conclusão do estudo vivo	Total
Tendência central				
Média	4,61	5,05	4,90	4,86
Mediana	4,50	4,85	4,80	4,70
Desvio padrão	1,33	1,36	2,14	1,23
Mínimo	1,06	2,20	2,10	1,06
Percentil 25	3,70	4,10	4,00	4,00
Percentil 75	5,30	5,10	5,60	5,60
Máximo	13,20	12,55	10,20	13,20
Dispersão interna				
Média	28,44	27,47	27,62	27,78
Mediana	24,46	23,96	25,00	24,74
Desvio padrão	19,08	17,39	16,07	17,29
Mínimo	0,00	0,00	0,00	0,00
Percentil 25	16,13	16,28	17,73	17,14
Percentil 75	36,67	34,89	33,44	34,15
Máximo	133,33	121,05	342,57	342,57
Tendência temporal				
Média	-0,03	-0,02	-0,01	-0,01
Mediana	0,00	0,00	0,00	0,00
Desvio padrão	0,43	0,53	0,36	0,40
Mínimo	-2,45	-3,50	-2,85	-3,50
Percentil 25	0,10	0,10	0,10	0,10
Percentil 75	0,10	0,10	0,10	0,10
Máximo	2,55	3,00	2,30	3,00

com desvios padrão elevados (em torno de 16 a 19) e valores extremos que chegam a 342 mg/dL. Esses registros refletem tanto a presença de dados atípicos quanto a instabilidade do metabolismo mineral em pacientes renais. O grupo de óbito apresenta ligeiramente maior variabilidade, sugerindo que oscilações mais intensas nos níveis de fósforo podem estar relacionadas a pior prognóstico.

Na dimensão de tendência temporal, a mediana das diferenças sucessivas é nula em todos os grupos, indicando ausência de tendência sistemática de aumento ou redução ao longo do tempo. Entretanto, os valores mínimos negativos (até –3,5 mg/dL) e máximos positivos (até +3,0 mg/dL) revelam oscilações pontuais, que podem refletir tanto ajustes terapêuticos quanto episódios de descontrole metabólico.

Em síntese, os resultados mostram que, embora os níveis médios de fósforo se mantenham próximos entre os grupos, a variabilidade interna e as oscilações temporais desempenham papel relevante na caracterização do risco. Isso justifica a inclusão do fósforo como covariável na análise de sobrevivência, dada sua importância clínica para complicações cardiovasculares e progressão da DRC.

5.5 Modelo de Cox preliminar para avaliação de covariáveis

O modelo de riscos proporcionais de Cox foi adotado nesta etapa como análise preliminar por ser amplamente utilizado em estudos de sobrevivência. Ele permite avaliar simultaneamente o impacto das covariáveis sobre o tempo até o evento sem necessidade de especificar a forma da função de risco de base, focando apenas no efeito relativo das variáveis.

Neste modelo inicial foram incluídas todas as variáveis clínicas, laboratoriais e demográficas selecionadas. O objetivo é verificar se o conjunto contribui para explicar a ocorrência do evento e identificar quais covariáveis apresentam significância estatística, funcionando como filtro analítico para reduzir a dimensionalidade do banco de dados.

O modelo fornece estimativas dos coeficientes e das razões de risco, que indicam a direção e magnitude dos efeitos: valores acima de 1 representam aumento do risco e abaixo de 1 indicam efeito protetor. Essa etapa preliminar estabelece uma base sólida para a aplicação posterior de técnicas de aprendizado de máquina, que poderão explorar interações complexas e melhorar a capacidade preditiva.

O modelo de Cox foi ajustado com 4.342 pacientes, dos quais 938 (21,6%) apresentaram o evento e 3.404 (78,4%) foram censurados. Não houve exclusão de casos por dados ausentes ou inconsistências, garantindo a integridade da amostra. O teste global do modelo indicou significância global do modelo (Qui-quadrado = 883,2; p $<$ 0,001), confirmando que o conjunto de covariáveis contribui para explicar o risco de ocorrência do evento.

Para a interpretação dos resultados, considerou-se como critério de significância estatística o valor de $p < 0,05$. Dessa forma, apenas as covariáveis com valores de p inferiores a este limite foram analisadas quanto ao risco associado.

Entre as covariáveis analisadas, destacaram-se como **fatores de risco** a idade, em que cada ano adicional aumenta em aproximadamente 2,3% a probabilidade de ocorrência do evento, e a tendência temporal da hemoglobina, cujas reduções sucessivas se associam a um acréscimo de 27,5% no risco. Esses resultados reforçam que tanto o envelhecimento quanto a instabilidade hematológica constituem elementos centrais na determinação do prognóstico dos pacientes.

Por outro lado, foram identificados como **fatores protetores** o escore de desempenho funcional (Karnofsky), associado a uma redução de 3,6% no risco por ponto

adicional; níveis mais elevados de pressão arterial sistólica, com 1,3% menos risco por unidade; hemoglobina, com 11,5% menos risco por unidade; e potássio, que reduz em 20,1% o risco por unidade adicional. Também se destacaram a estabilidade temporal da pressão arterial, com 3,4% menos risco por unidade, e do potássio, com 28,0% menos risco por unidade. Em conjunto, esses achados evidenciam que melhores condições clínicas e laboratoriais, associadas à manutenção de parâmetros estáveis ao longo do tempo, reduzem significativamente a probabilidade de ocorrência do evento. Assim, o modelo preliminar de Cox permite distinguir variáveis que aumentam ou reduzem o risco, fornecendo uma base sólida para análises preditivas posteriores.

A Tabela 32 apresenta os resultados do modelo de Cox para todas as covariáveis consideradas.

Tabela 32 – Resultados completos do modelo de Cox preliminar

Covariável	**Coeficiente (B)**	**R. de risco (Exp(B))**	**Valor-p**	**Significativa**
Branco	0,161	1,175	0,417	Não
Preto	0,170	1,186	0,439	Não
Pardo	0,229	1,257	0,278	Não
Até 2 s.m.	0,452	1,572	0,209	Não
Entre 2 e 5 s.m.	0,319	1,376	0,376	Não
5 ou mais s.m.	0,403	1,496	0,268	Não
Idade	0,023	1,023	$< 0,001$	Sim
Karnofsky	-0,036	0,964	$< 0,001$	Sim
Doença Cardiovascular	0,072	1,074	0,346	Não
Diabetes	0,108	1,114	0,120	Não
HAS	0,053	1,055	0,571	Não
PAS (tendência Central)	-0,013	0,987	$< 0,001$	Sim
HEM (tendência Central)	-0,123	0,885	$< 0,001$	Sim
FOS (tendência Central)	0,004	1,004	0,907	Não
POT (tendência Central)	-0,225	0,799	$< 0,001$	Sim
PAS (Dispersão interna)	0,001	1,001	0,647	Não
HEM (Dispersão interna)	0,032	1,033	0,281	Não
FOS (Dispersão interna)	-0,009	0,991	0,857	Não
POT (Dispersão interna)	-0,012	0,988	0,800	Não
PAS (Tendência temporal)	-0,035	0,966	$< 0,001$	Sim
HEM (Tendência temporal)	0,243	1,275	0,002	Sim
FOS (Tendência temporal)	-0,052	0,949	0,654	Não
POT (Tendência temporal)	-0,328	0,720	0,014	Sim

Os resultados confirmam a relevância da idade, estado funcional e de variáveis

clínicas e laboratoriais como pressão arterial sistólica, hemoglobina e potássio na determinação do risco de evento. Além dos valores centrais, as tendências temporais mostraram-se importantes, evidenciando que oscilações longitudinais carregam informação prognóstica adicional. Esse modelo preliminar permite reduzir o banco de dados às covariáveis mais relevantes, servindo como base para a etapa seguinte de aplicação de algoritmos de aprendizado de máquina, que poderão explorar interações complexas e melhorar a capacidade preditiva.

Entre os achados específicos, observou-se que níveis sustentados mais altos de hemoglobina, pressão arterial sistólica e potássio estão associados a menor risco de evento, enquanto a variabilidade das medidas não apresentou significância estatística. Já a tendência temporal da hemoglobina mostrou-se paradoxalmente associada a maior risco, possivelmente refletindo gravidade clínica ou efeito de confusão. Em contraste, as tendências temporais da pressão arterial sistólica e do potássio reforçaram o efeito protetor, sugerindo que a estabilidade longitudinal desses parâmetros desempenha papel relevante no prognóstico.

Como conclusão, o modelo de Cox indica que medidas sustentadas e estáveis de pressão arterial, hemoglobina e potássio são protetoras, enquanto quedas longitudinais da hemoglobina podem sinalizar maior risco. Esse contraste entre efeitos centrais e temporais sugere que a suposição de riscos proporcionais pode não ser plenamente atendida, apontando para a possível presença de relações não lineares ou interações complexas entre covariáveis. Tal padrão justifica a adoção de técnicas mais flexíveis, como modelos dependentes do tempo ou algoritmos de aprendizado de máquina, na etapa seguinte da análise.

Sobre as variáveis que não apareceram como significativas no modelo de Cox, algumas circunstâncias específicas do tratamento dos pacientes podem gerar hipóteses para essa falta de significância:

- No caso das variáveis demográficas, como raça e renda, é possível que o funcionamento do SUS tenha reduzido diferenças de acesso ao tratamento. Como o atendimento é custoso mas garantido pelo sistema público, essas variáveis podem perder força estatística, embora em contextos privados elas costumem influenciar bastante.

- Quanto às comorbidades, uma suspeita é que os pacientes já se encontram em estágio avançado da doença renal crônica, pois estão em diálise peritoneal. Nesse cenário, o impacto adicional de doenças associadas pode ser menor, já que a condição principal domina o risco de sobrevida.

- Em relação ao fósforo, mesmo sendo um marcador importante na DRC, o acompanhamento contínuo e abrangente realizado pelo SUS pode estar controlando seus níveis de forma sistemática. Isso reduziria a variabilidade e, consequentemente, o efeito estatístico no modelo.

5.6 Aplicação de modelo adaptativo sobre banco reduzido

Nesta seção apresentam-se os resultados da aplicação de cinco modelos sobre um conjunto de dados de pacientes com DRC, considerando apenas as covariáveis significativas identificadas na seção anterior. O desempenho foi avaliado por meio das métricas **C-Index** e **IBS**, amplamente utilizadas em análises de sobrevivência. Para a avaliação deste conjunto de dados, empregou-se o mesmo procedimento descrito na Subseção 4.3.2.

Um primeiro resultado relevante é que, nas 20 repetições realizadas, o modelo adaptativo selecionou sempre a função de ativação sigmoide. Esse comportamento sugere que uma estrutura não linear é a mais adequada para analisar este conjunto de dados. Além disso, o valor médio dos graus de liberdade da distribuição *t-Student* foi de aproximadamente 16,5, indicando um tratamento consistente de dados extremos, o que contribui para melhorar as previsões.

A Tabela 33 apresenta os valores médios e intervalos de confiança para C-Index e IBS obtidos ao longo das 20 repetições.

Tabela 33 – DRC Reduzido: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
CPH	0.749 (0.740,0.758)	0.213 (0.153,0.273)
BJ-LS	0.747 (0.739,0.754)	0.269 (0.171,0.367)
BJ-ELM	0.792 (0.784,0.800)	0.212 (0.159,0.264)
BJ-ELML (Adaptativo)	0.792 (0.784,0.800)	0.212 (0.159,0.264)
BJ-ELMLR (Adaptativo)	0.792 (0.784,0.799)	0.210 (0.159,0.261)

A tabela mostra que o CPH obteve C-Index de 0.749 e IBS de 0.213, confirmando seu desempenho sólido, mas limitado frente a modelos mais modernos. O BJ-LS apresentou resultados semelhantes em C-Index (0.747), porém com IBS mais elevado (0.269), indicando pior calibração e menor precisão preditiva.

Os modelos baseados em BJ-ELM destacaram-se: alcançaram C-Index de 0.792, superior aos métodos tradicionais, e IBS em torno de 0.212, mostrando melhor equilíbrio entre discriminação e calibração. O BJ-ELML adaptativo manteve desempenho idêntico ao BJ-ELM, evidenciando estabilidade. Já o BJ-ELMLR adaptativo robusto obteve o mesmo C-Index, mas reduziu ligeiramente o IBS para 0.210, sugerindo maior robustez frente a dados extremos.

É importante destacar que, para o C-Index, os intervalos de confiança dos modelos BJ-ELM, BJ-ELML e BJ-ELMLR não se sobrepõem com os intervalos de CPH e BJ-LS, reforçando a evidência de desempenho significativamente superior. No caso do IBS, embora os valores médios dos modelos BJ-ELM sejam menores, há sobreposição parcial dos intervalos de confiança, o que indica uma diferença menos pronunciada, mas ainda favorável às versões adaptativas e robustas.

A Figura 18 apresenta os box-plots das distribuições de C-Index e IBS obtidas nas 20 repetições para os modelos avaliados, permitindo visualizar a variabilidade dos resultados e a presença de dados extremos.

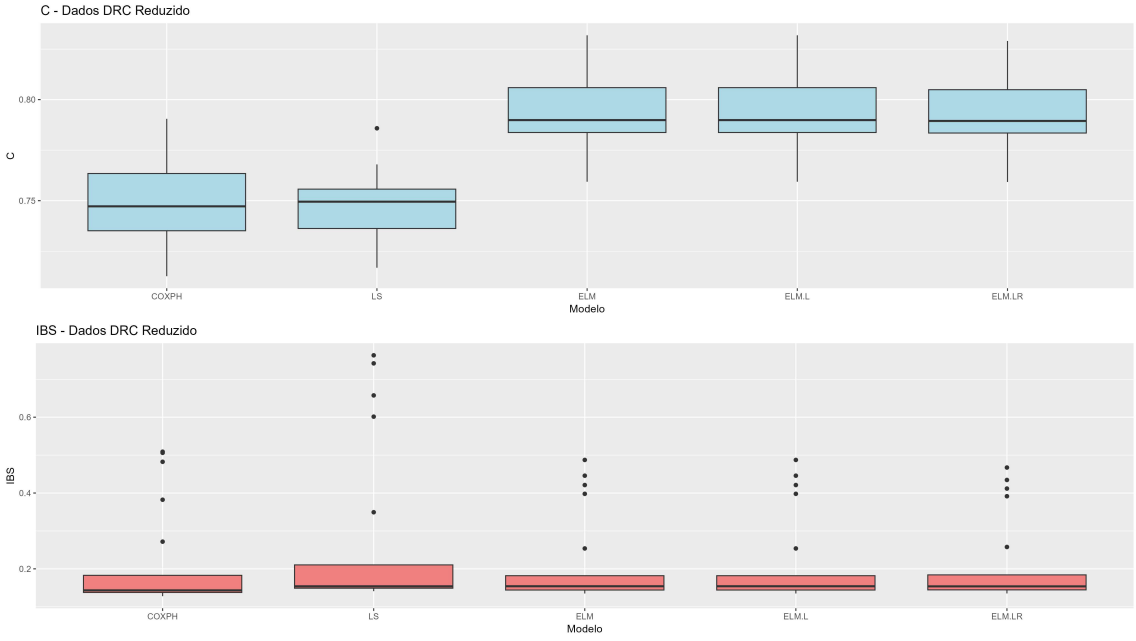

Figura 18 – DRC Reduzido: Box-Plot para C-Index e IBS para modelos avaliados.

A análise dos gráficos evidencia que os modelos BJ-ELM e suas variantes adaptativas apresentam desempenho superior aos métodos tradicionais (CPH e BJ-LS), com maior capacidade discriminativa e melhor calibração. Entre os modelos adaptativos, o BJ-ELMLR se destaca como o mais promissor, por combinar alto poder preditivo com maior robustez, tornando-se uma alternativa competitiva aos demais modelos avaliados. Essas conclusões são consistentes com os resultados apresentados nas tabelas, reforçando a superioridade dos modelos baseados em BJ-ELM frente às abordagens tradicionais.

5.7 Aplicação de modelo adaptativo sobre banco completo

Como avaliação final, nesta seção apresentam-se os resultados da aplicação de cinco modelos sobre um conjunto de dados de pacientes com DRC, considerando todas as variáveis utilizadas no modelo Cox preliminar. O objetivo é verificar se alguma covariável previamente excluída pode ser relevante por apresentar uma relação não linear não observada, contribuindo para aprimorar o modelo reduzido. O desempenho foi avaliado nos mesmos moldes da seção anterior, seguindo o procedimento descrito na Subseção 4.3.2.

Nesta avaliação, o modelo adaptativo selecionou sempre a função de ativação sigmoide, sugerindo novamente que a estrutura não linear é adequada. No entanto, o valor médio dos graus de liberdade da distribuição *t-Student* foi de aproximadamente 30, indicando que não é necessário o tratamento de dados extremos.

A Tabela 34 apresenta os valores médios e intervalos de confiança para as métricas C-Index e IBS, obtidos ao longo das 20 repetições no conjunto completo de dados de pacientes com DRC. Observa-se que o modelo CPH manteve desempenho sólido (C-Index = 0.746; IBS = 0.215), enquanto o BJ-LS apresentou resultados inferiores (C-Index = 0.721; IBS = 0.255). Os modelos baseados em BJ-ELM e suas variantes adaptativas alcançaram C-Index em torno de 0.781 e IBS próximo de 0.219–0.221, indicando desempenho superior ao BJ-LS, mas sem vantagem clara em relação ao CPH.

Tabela 34 – DRC Completo: Resultados para C-Index e IBS em modelos avaliados.

Modelo	$C-INDEX$	IBS
CPH	0.746 (0.737,0.755)	0.215 (0.154,0.276)
BJ-LS	0.721 (0.713,0.728)	0.255 (0.181,0.328)
BJ-ELM	0.781 (0.775,0.788)	0.219 (0.162,0.275)
BJ-ELML (Adaptativo)	0.781 (0.775,0.788)	0.219 (0.162,0.275)
BJ-ELMLR (Adaptativo)	0.780 (0.774,0.787)	0.221 (0.164,0.278)

Quando comparados aos resultados obtidos no conjunto reduzido, nota-se que os modelos BJ-ELM e variantes adaptativas apresentaram desempenho inferior no conjunto completo: o C-Index caiu de 0.792 para aproximadamente 0.781 e o IBS aumentou de cerca de 0.210–0.212 para 0.219–0.221. Essa diferença sugere que a inclusão de todas as covariáveis não trouxe benefício adicional e pode ter introduzido problemas como multicolinearidade ou ruído, reduzindo a capacidade discriminativa e a calibração dos modelos. Assim, o conjunto reduzido mostrou-se mais eficiente e robusto para a análise de sobrevivência.

A Figura 19 apresenta os box-plots das distribuições de C-Index e IBS obtidas nas 20 repetições para os cinco modelos avaliados no conjunto completo de dados de pacientes com DRC. Observa-se que os modelos BJ-ELM e suas variantes adaptativas (BJ-ELML e BJ-ELMLR) apresentam as maiores medianas de C-Index, indicando melhor capacidade discriminativa. No entanto, esses modelos exibem maior dispersão e presença de *outliers* em algumas repetições, sugerindo instabilidade relativa. O modelo CPH mantém desempenho sólido, com mediana intermediária e menor variabilidade, enquanto o BJ-LS apresenta os piores resultados, com C-Index mais baixo e IBS mais elevado. Para o IBS, os modelos BJ-ELM e variantes adaptativas mostram valores medianos próximos aos do CPH, mas com maior variabilidade, ao passo que o BJ-LS novamente se destaca negativamente por apresentar os maiores valores.

Ao comparar com os resultados obtidos no conjunto reduzido, nota-se que os modelos avaliados no conjunto completo apresentam maior dispersão e maior presença de *outliers*, especialmente nos modelos baseados em EML. No conjunto reduzido, os modelos BJ-ELM e variantes adaptativas mostraram desempenho mais estável, com menor variabilidade e melhor calibração (IBS mais baixo), além de C-Index superior. Isso reforça

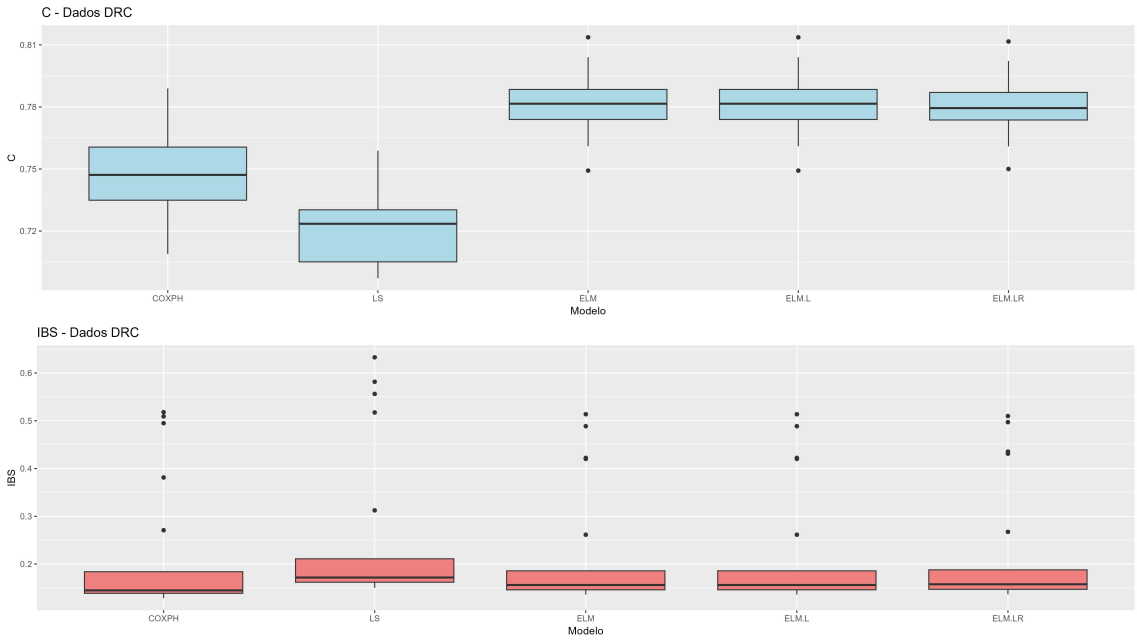

Figura 19 – DRC Completo: Box-Plot para C-Index e IBS para modelos avaliados.

a hipótese de que a inclusão de variáveis irrelevantes ou colineares no conjunto completo pode ter introduzido ruído e instabilidade, prejudicando a robustez dos modelos. Assim, o conjunto reduzido se mostra mais eficiente e confiável para análise de sobrevivência. Dessa forma, os resultados evidenciam a importância de se realizar uma etapa de análise preliminar, a fim de orientar a seleção adequada das covariáveis e indicar o caminho mais consistente para a conclusão das análises.

6 CONCLUSÕES DO TRABALHO E PROPOSTAS FUTURAS

O presente trabalho de doutorado propôs e avaliou um algoritmo adaptativo para modelagem de dados de sobrevivência com censura à direita, capaz de lidar de forma eficiente com relações complexas (iterações e não linearidade) e presença de dados extremos (*outliers*). O algoritmo mostrou-se flexível e adaptativo às características específicas dos conjuntos de dados analisados. O teste RESET foi empregado como ferramenta de diagnóstico para determinar a linearidade do modelo e orientar a escolha da função de ativação no ELM, confirmando a adequação dos diagnósticos realizados. A proposta, fundamentada no modelo de Buckley-James (BJ), demonstrou maior consistência em comparação ao modelo de Cox, ao contornar a suposição de riscos proporcionais. Além disso, a utilização de um comitê de Máquinas de Aprendizado Extremo (ELM) com estratégia de Boosting L2 conferiu a flexibilidade necessária para tratar relações complexas, enquanto a adoção da distribuição *t-Student* para os erros aumentou a robustez do modelo.

A aplicação do algoritmo ao conjunto de dados de pacientes com doença renal crônica em tratamento por diálise peritoneal mostrou-se bem-sucedida. O processo de análise evidenciou a importância das etapas de pré-processamento e da definição de um roteiro analítico. Nessas condições, o algoritmo proposto mostrou-se a melhor opção para a previsão dos tempos de sobrevida. O indicativo de não linearidade sugerido na análise preliminar, via modelo de Cox, foi corroborado pela aplicação do modelo desenvolvido, cuja superioridade pôde ser verificada pela medida de desempenho C-Index, com intervalos de confiança substancialmente superiores aos modelos lineares, e pela medida IBS, que apresentou o menor valor.

Em síntese, as conclusões confirmam o atendimento ao objetivo proposto, demonstrando que o algoritmo BJ-ELMLR Adaptativo constitui uma alternativa robusta e eficiente para análise de dados de sobrevivência em cenários complexos. Além de validar sua aplicação em pacientes renais, os resultados reforçam o potencial da metodologia para estudos futuros em diferentes contextos clínicos e epidemiológicos.

Com base nos resultados alcançados neste trabalho, algumas conclusões específicas podem ser destacadas:

- No geral, os resultados das simulações realizadas mostram que o método proposto é mas eficiente e consistente comparado ao modelo de riscos proporcionais de Cox, apresentando uma melhor adequação às diversas características dos dados;
- A escolha do teste RESET para avaliação da linearidade do modelo mostrou-se eficiente. O resultado do teste determina qual função de ativação será usada no ELM, considerando um nível de significância de 5%. Assim, opta-se entre a função identidade (modelo linear) ou a função sigmoide (modelo não linear);

- A escolha da distribuição *t-Student* como alternativa à normalidade dos erros influenciou positivamente o processo de estimação dos parâmetros, melhorando as predições segundo as medidas de desempenho avaliadas. Observou-se que o grau de liberdade varia conforme o conjunto de dados, indicando que o modelo se adapta às características específicas de cada situação;
- A flexibilidade proporcionada pela escolha de diferentes valores do grau de liberdade da distribuição *t-Student*, permite que o método ajuste uma ponderação adequada das observações no processo de estimação. O ajuste dinâmico desse grau de liberdade em cada iteração contribui para resultados mais consistentes;
- A possibilidade de atualização automática, tanto na escolha da função de ativação via teste RESET quanto na definição do grau de liberdade da distribuição *t-Student*, reforça a capacidade adaptativa do modelo proposto às características específicas dos conjuntos de dados analisados;
- Nos testes iniciais com dados sintéticos da literatura, considerando o desempenho avaliado pelas medidas C-Index e IBS, observa-se que a proposta apresentada oferece melhora na predição dos tempos de sobrevida. Uma exceção ocorre quando as covariáveis não são correlacionadas, condição pouco frequente em situações reais. Mesmo nesses casos, o modelo proposto aproxima-se dos modelos lineares mais conhecidos;
- Na análise dos dados de pacientes com doença renal crônica em tratamento por diálise peritoneal, o algoritmo proposto demonstrou desempenho superior em relação aos modelos lineares tradicionais. Os resultados evidenciaram intervalos de confiança mais elevados para o C-Index e valores menores para o IBS, confirmando a capacidade do método em capturar a não linearidade presente nos dados clínicos e em fornecer predições mais precisas dos tempos de sobrevida.

Considerando os resultados obtidos nesta tese, destacam-se algumas possibilidades de investigação a serem exploradas em trabalhos futuros:

- Incluir dentro do processo adaptativo critérios de regularização do modelo, como a inclusão de Lasso, Ridge ou, de forma mais geral, Elastic Net. A regularização contribui para reduzir sobreajuste, melhorar a estabilidade das estimativas e realizar seleção automática de variáveis, sendo especialmente útil em contextos de alta dimensionalidade;
- Incluir um fator de correção que controle a assimetria na variável resposta. No modelo Buckley-James, a transformação logarítmica do tempo de sobrevida proporciona maior simetria em relação aos dados originais, mas pode ainda apresentar certa assimetria;

- Avaliar a possibilidade de utilizar os pesos da camada oculta e da camada de saída do ELM para interpretar de forma descritiva os efeitos das covariáveis consideradas;
- Considerar melhorias no uso do teste RESET para determinação da linearidade ou não do modelo. Do ponto de vista da inferência estatística, o poder do teste depende do número de covariáveis e do tamanho da amostra. Assim, pode-se pensar em ajustes que levem em conta essa relação, tornando o teste mais eficiente em diferentes situações;
- Explorar o uso de modelos paramétricos de análise de sobrevivência, aproveitando a estrutura do algoritmo proposto para comparar desempenho e robustez em diferentes cenários;
- Generalizar a proposta do algoritmo para modelos com dados censurados, ampliando sua aplicação além dos modelos de sobrevivência tradicionais. Dessa forma, o método poderá ser utilizado em diferentes contextos estatísticos em que a censura esteja presente.

REFERÊNCIAS

[1] KENNETH, J. R. The rise and fall of epidemiology, 1950-2000 A.D. *International Journal of Epidemiology*, v. 36, n. 4, p. 708-710, 2007.

[2] FORFAR, D.O. Mortality Laws. In: *Encyclopedia of Actuarial Science*, v2, p 1-6, Wiley, 2006.

[3] Smoking and Health: report of the Advisory Committee to the Surgeon General of the Public Health Service, Washington, D.C.: Government Printing Office, 1964. (DHEW publication no. (PHS) 1103).

[4] ESCOBAR, R.; VILLA, E.; YAÑEZ, S. Confiabilidad: Historia, estado del arte y desafios futuros. *Dyna*, v. 70, n. 140, p. 5-21, 2003.

[5] AZARKHAIL, M.; MODARRES, M. The Evolution and History of Reliability Engineering: Rise of Mechanistic Reliability Modeling. *International Journal of Performability Engineering*, v. 8, n. 1, p. 35-47, 2012.

[6] DIRICK, L.; CLAESKENS, G. ; BAESENS, B. Time to default in credit scoring using survival analysis: a benchmark study. **Journal of Operacional Research Society**, v. 68, p. 652-665, 2017.

[7] WANG, P.; LI, Y.; REDDY, C. Machine Learning for Survival Analysis: A Survey. **ACM Computing Surveys**, v. 51, n. 6, p. 1-36, 2019.

[8] ISHWARAN, H.; KOGALUR, U. Random Survival Forests for High-Dimensional Data. **Statistical Analysis and Data Mining: The ASA Data Science Journal**, 4.1 (2011): 115-132, 2011.

[9] ZHAO, W.; LUO, J.; CAO, Z. Deep Learning for High-dimensional Data in Survival Analysis: An Overview. **Statistical Methods in Medical Research**, 28(12), 3481-3495, 2019.

[10] CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 785-794, 2016.

[11] ZHU R.; ZENG D.; KOSOROK M. C. Boosting for High-Dimensional Sparse Survival Data, with an Application to Prediction of Alzheimer's Disease Progression. **Biostatistics**, 18(4), 605-619.

[12] HUANG, G.; ZHU, Q.; CHEE KHEONG, S. Extreme learning machine: a new learning scheme of feedforward neural networks. **In Neural Networks, 2004. Proceedings. 2004 IEEE Inter national Joint Conference**, volume 2, pages 985–990. IEEE, 2004.

[13] HUANG, G.; ZHU, Q.; CHEE KHEONG, S. Extreme learning machine: Theory and applications. **Neurocomputing**, 70.1-3, pp 489-501, 2006.

[14] WANG, J.; LU, S; WANG, SH; ZHANG YD. A review on extreme learning machine. **Multimed Tools Appl**, 81, 41611–41660, 2022.

[15] WANG, H.; WANG, J.; ZHOU, L. A survival ensemble of extreme learning machine. **Applied Intelligence**, v. 48(4), 2018.

[16] WANG, H.; LI, G. Extreme learning machine Cox model for high-dimensional survival analysis. **Statistics in medicine**, Volume 38(12), p. 2139-2156, May 2019.

[17] KONG, J.; ZHANG, S. Buckley-James boosting model based on extreme learning machine and random survival forests. **Biometrical Jornal**, Jun;65(5):e2200153, 2023.

[18] COX, D.R. Regression Models and Life Tables (with discussion). **Journal Royal Statistical Society**, B, 34, p. 187-220, 1972.

[19] BUCKLEY, J.; JAMES, I. Linear regression with censored data. **Biometrika**, v. 66, n. 3, p. 429-436, 1979.

[20] MASSUIA, M.; CABRAL, C; MATOS, L; LACHOS, V. Influence diagnostics for student-t censored linear regression models. **Statistics: A Journal of Theoretical and Applied Statistics**, Vol. 49, no. 5, p. 1074-1094, 2014.

[21] KAPLAN, E.L.; MEIER, P. Nonparametric Estimation from Incomplete Observations. **Journal of the American Statistical Association**, v. 53, n. 282, p. 457-481, 1958.

[22] NELSON, W. Theory and applications of hazard plotting for censored failure data. **Technometrics**, 14(4), 945-966, 1972.

[23] AALEN, O. Nonparametric inference for a family of counting processes. **The Annals of Statistics**, 6(4), 701-726, 1978.

[24] CUTLER, SJ; EDERER, F. Maximum utilization of the life table method in analyzing survival. **Journal of Chronic Diseases**, 8(6), 699-712, 1958.

[25] COX, D.R. Partial Likelihood. **Biometrika**, v. 62, p. 269-276, 1975.

[26] KLEIN, JP.; MOESCHBERGER, ML. **Survival Analysis: Techniques for Censored and Truncated Data (2nd ed.**. Springer, 2003.

[27] THERNEAU, TM.; GRAMBSCH, PM. **Modeling Survival Data: Extending the Cox Model**. Springer, 2000.

[28] KALBFLEISCH, JD.; PRENTICE, RL. **The Statistical Analysis of Failure Time Data (2nd ed.)**. Wiley, 2002.

[29] BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J. **Classification and Regression Trees**. CA: Wadsworth International Group., 1984.

[30] BOU-HAMAD, I.; LAROCQUE, D.; BEN-AMEUR, H. A review of survival trees. **Statistics Surveys**, 5: 44–71, 2011.

[31] TIBSHIRANI, R. The Lasso Method for Variable Selection in the Cox Model. **Statistics in Medicine**, 16(4), 385-395, 1997.

[32] SIMON, N.; FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. **Journal of Statistical Software**, v. 39, p. 1-13, 2011.

[33] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.)**. Springer, 2009.

[34] WANG, H.; LI, G.; JIANG, G. Regularized Buckley–James method for high dimensional survival data. **BMC Bioinformatics**, 9(1), 1-13, 2008.

[35] LEE, ET.; WANG, JW. **Statistical Methods for Survival Data Analysis (3rd ed.)**. Wiley, 2002.

[36] ISHWARAN, H.; KOGALUR, UB. Random survival forests for R. **R News**, 7(2), 25-31, 2007.

[37] FARAGGI, D.; SIMON, R. Maximum Likelihood Neural Network Prediction Models. **Biometrical Journal**, Volume 37, issue 96, p. 713-725, 1995.

[38] KATZMAN, JL.; SHAHAM, U.; CLONINGER, A.; BATES, J.; JIANG, T.; KLUGER, Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. **BMC Medical Research Methodology**, 18(1), 24, 2018.

[39] VAN BELLE, V.; PELCKMANS, K.; VAN HUFFEL, S.; SUYKENS, JA. Support vector methods for survival analysis: A comparison between ranking and regression approaches. **Artificial Intelligence in Medicine**, 53(2), 107-118, 2011.

[40] ROTHMAN, K.; GREENLAND, S. *Modern Epidemiology*. Philadelphia: Lippincott-Raven, 1998.

[41] NELSON, W.; GREENLAND, S. *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*. New York: Wiley Series in Probability and Statistics, John Wiley Sons, 1990.

[42] COLOSIMO, E.; GIOLO, s. **Análise de Sobrevivência Aplicada**. São Paulo: Ed. Blucher, 2006.

[43] YAN, Y.; ZOU, H. A cocktail algorithm for solving the elastic net penalized Cox regression in high dimensions. **Statistics and Its Interface**, v. 6, p. 167-173, 2013.

[44] TIBSHIRANI, R. What is Cox's proportional hazards model?. **Significance**, v. 19, n. 2, p. 38-39, 2022.

[45] FIRTH, D.; REID, N.; MAYO, D.G.; BATTEY, H. Remembering Sir David Cox, 1924–2022. **Significance**, v. 19, n. 2, p. 30-37, 2022.

[46] R CODE TEAM. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. URL https://www.R-project.org/, 2021.

[47] HARRELL, FE.; LEE, KL.; MARK, DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. **Statistics in Medicine**, 15(4), 361-387, 1996.

[48] GRAF, E.; SAUERBREI, W.; SCHUMACHER, M. Assessment and comparison of prognostic classification schemes for survival data. **Statistics in Medicine**, 18(17-18), 2529-2545, 1999.

[49] FRIEDMAN, J. Greedy Function Approximation: A Gradient Boosting Machine. **Annals of Statistics**, 29(5), 1189-1232, 2001.

[50] WANG, Z. WANG, C., Buckley-James boosting for survival analysis with high-dimensional biomarker data. **Statistical Applications in Genetics and Molecular Biology**, 9, Article 24, 2010.

[51] SUASSUNA, N.M., Diálise Peritoneal no Brasil: Descrição de uma Coorte e Fatores de Risco para Sobrevivência da Técnica e do Paciente. **Tese de Doutorado Programa de Pós Graduação em Saúde**, Universidade Federal de Juiz de Fora, 2009.

[52] CURIOSO, I. SANTOSA, R. . RIBEIRO, B. CARREIRO, A. COELHO, P. FRAGATA, J . GAMBOA, H., Addressing the Curse of Missing Data in Clinical Contexts: A Novel Approach to Correlation-based Imputation. **Journal of King Saud University– Computer and Information Sciences**, 35 101562, 2023.

[53] WILCOX, R. **Introduction to robust estimation and hypothesis testing**. 3. ed. Amsterdam: Academic Press, 2012.

[54] TUKEY, J. **Exploratory data analysis**. Reading, MA: Addison-Wesley, 1977.

[55] KARNOFSKY DA, ABELMANN WH, CRAVER LF, BURCHENAL JH **The Clinical Evaluation of Chemotherapeutic Agents in Cancer**. In: MacLeod CM (ed.). Evaluation of Chemotherapeutic Agents, edited by Mc Leod CM, New York Columbia University Press, p. 191-205, 1949.

[56] ANDREWS,DF.;MALLOWS SL, Scale mistures of normal distributions. **Journal of the Royal Statistical Society**, Series B, 36, 99-102, 1974.

[57] HUAIRA CONTRERAS, CA., Modelo de regressão linear mistura de escala normal com ponto de mudança: Estimação e diagnóstico. **Dissertação de mestrado**, Universidade Estadual de Campinas, 2014.

[58] RAMSEY,JB, Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. **Journal of the Royal Statistical Society**, Series B, 31(2), 350–371, 1969.

[59] ZEILEIS,A, HOTHORN, T. , Diagnostic Checking in Regression Relationships. **R News**, 2(3), 7–10, 2002.

[60] CASELLA, G.; BERGER, RL. **Inferência Estatística**. 2. ed. São Paulo: Cengage Learning, 2010.

[61] CHEN,K; LV q.; LU Y.; DOU Y.Robust regularized extreme learning machine for regression using iteratively reweighted least squares. **Neurocomputing**, 230, 489–501, 2017.