Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Imputação de dados no modelo de regressão de Cox na presença da verossimilhança monótona

por

João Pedro de Oliveira Morais

Brasília, 18 de julho de 2025

Imputação de dados no modelo de regressão de Cox na presença da verossimilhança monótona

por

João Pedro de Oliveira Morais

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Mestre em Estatística

Orientador: Prof. Dr. Frederico M. Almeida

Brasília, 18 de julho de 2025

Dissertação submetida ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do título de Mestre em Estatística.

Texto aprovado por:

Prof. Dr. Frederico Machado Almeida
Orientador, EST/UnB

Prof. Dr. Eduardo Yoshio Nakano
Membro interno, EST/UnB

Prof. Dr. Fábio Prataviera
Membro externo, ESALQ/USP

Prof. Dr. Helton Saulo Bezerra dos Santos
Suplente, EST/UnB

A vida não é sobre quão duro você bate, mas sobre quão duro você pode apanhar e continuar seguindo em frente.

(Rocky Balboa)

Dedico este trabalho à minha amada esposa Déborah Santos e à minha querida família.

Agradecimentos

Agradecimento a Deus que me deu forças para superar os desafios do mestrado. Agradecimentos ao PPGE/UnB e ao Prof. Dr. Frederico Machado Almeida, meu orientador.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

O problema envolvendo dados faltantes, tem sido recorrente em diversas áreas de pesquisa, e em estudos de análise de sobrevivência não é diferente. Na literatura, a abordagem que tem sido considerada com alguma frequência tem sido a remoção de todas as observações com pelo menos uma célula faltante nas covariáveis. Entretanto, tal procedimento pode, de alguma forma acarretar a perda de informação. Além disso, descartar observações faltantes pode igualmente introduzir um viés nas estimativas dos parâmetros, o que pode ocasionar uma certa instabilidade nos procedimentos de estimação, como por exemplo, a ocorrência do problema da verossimilhança monótona. A motivação deste trabalho é o banco de dados de melanoma contendo mais de 50% de observações faltantes. Além disso, a eliminação de tais observações acarretou uma instabilidade nos procedimentos de estimação (não existência de estimativas finitas para os coeficientes de regressão). Atendendo o exposto anteriormente, este trabalho propõe os métodos de imputação de dados como forma de não perder informação relevante dos dados, e ao mesmo tempo, reduzir as chances de ocorrência da verossimilhança monótona. Para aferir o desempenho do modelo em termos de estimação, um estudo de simulações é proposto avaliando diferentes taxas de dados faltantes, e percentagens de amostras com verossimilhança monótona.

Palavras-chave: análise de sobrevivência, dados faltantes, imputação de dados, verossimilhança monótona, melanoma.

Abstract

Imputation procedures for censored data in the presence of monotone likelihood

The problem of missing data has been recurrent in various research fields, and survival analysis studies are no exception. In the literature, a frequently employed strategy involves excluding all units with at least one missing value in covariates. However, such a procedure can lead to the loss of information. Moreover, discarding missing observations can also introduce bias in parameter estimation, which may result in some instability in the estimation procedures, such as the occurrence of the monotone likelihood problem. The motivation for this study comes from a melanoma dataset containing more than 50% missing observations. Furthermore, the removal of such observations resulted in instability in the estimation procedures, including the non-existence of finite estimates for the regression coefficients. In light of these issues, this study proposes the use of data imputation methods as a means to retain relevant information while simultaneously reducing the likelihood of monotone likelihood problems. To evaluate the model's performance in terms of estimation, a simulation study is conducted, examining varying levels of missing data and the proportion of samples exhibiting monotone likelihood. Finally, a real data application is presented that includes both the monotone likelihood issue and missing data.

Keywords: survival analysis, missing data, data imputation, monotone likelihood, melanoma.

Lista de Abreviaturas

ACM	Acurácia Média de Imputação
DATASUS	Departamento de Informática do Sistema Único de Saúde
dpAC	Desvio Padrão da Acurácia
dp(rEQM)	Desvio Padrão da Raiz do Erro Quadrático Médio
EAM	Erro Absoluto Médio
eMC	Estimativas de Monte Carlo
EM	Expectation-Maximization (Esperança-Maximização)
EMV	Estimador de Máxima Verossimilhança
EP	Erro Padrão
epMC	Erro Padrão de Monte Carlo
FCS	Fully Conditional Specification
IC	Intervalo de Confiança
IPW	Inverse Probability Weighting (Ponderação pela Probabilidade Inversa)
MAR	Missing at Random (Ausência Aleatória)
MCAR	Missing Completely at Random (Ausência Totalmente Aleatória)
MCMC	Cadeias de Markov Monte Carlo (Markov Chain Monte Carlo)
MICE	Multiple Imputation by Chained Equations
MNAR	Missing Not at Random (Ausência Não Aleatória)
NA	Not Available (Valor Ausente)

ONCAD	Oncologia Cirúrgica do Aparelho Digestivo
pCob	Probabilidade de Cobertura Empírica
rEQM	Raiz do Erro Quadrático Médio
RR	Razão de Risco
SMC-FCS	Substantive Model Compatible Fully Conditional Specification
UFMG	Universidade Federal de Minas Gerais
VM	Verossimilhança Monótona
VRM	Viés Relativo Médio

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

O modelo de regressão de Cox (Cox, 1972) tem sido amplamente utilizado para estudar a associação entre o tempo de sobrevivência e um conjunto de fatores de riscos, comumente designados por covariáveis. Sua versatilidade, associada com a suposição de riscos proporcionais estão entre os aspectos que propiciam a popularidade desse modelo em estudos da área de saúde (Gui e Li, 2005; Cherobin et al., 2018; Hughey et al., 2019; Moncada-Torres et al., 2021). É importante reforçar que a popularidade do modelo de regressão de Cox não se restringe apenas a pesquisas biomédicas, por exemplo, em demografia podemos citar os seguintes trabalhos (Andersen et al., 1985; Ihwah, 2015; Ayele et al., 2017), na engenharia (Wong, 2011; Ezzeddine et al., 2015; Thijssens e Verhagen, 2020), em criminologia (Barton e Turnbull, 1979; Kingsley et al., 2008; Tollenaar e Van Der Heijden, 2019), na área de seguro (Keiding et al., 1998; Sari et al., 2019), em ciências sociais (Kavkler et al., 2009; Abd ElHafeez et al., 2021), entre outras aplicações.

Um problema que pode ser considerado um desafio na modelagem estatística, é a presença de observações incompletas em determinadas características do conjunto de dados. Tais observações são comumente referenciadas por *dados ausentes* (ou *missing data*, sua designação do inglês). Dados ausentes (ou simplesmente valores faltantes) são definidos como valores ou observações que não são armazenados para as características de determinados indivíduos. Po-

dendo ser observado em pesquisas de diversas áreas,a ausência de tais observações pode resultar em uma série de problemas na modelagem, entre os quais se destacam: (i) a redução do poder dos testes estatísticos (podendo afetar a probabilidade dos erros tipo I e II), (ii) pode acarretar um viés no processo de estimação, (iii) redução do poder amostral (em termos de representatividade), e (iv) pode afetar a validade do modelo, e influenciar negativamente as conclusões que podem ser tiradas dos dados (Kang, 2013).

Os dados faltantes podem ocorrer por diversos motivos. Por exemplo, devido a problemas no processo de coleta, como erros de digitação ou falhas no registro de informações. Além disso, instrumentos de medição defeituosos ou mal calibrados, bem como falhas técnicas em software ou hardware durante a coleta de dados, também podem contribuir para a perda de informações. Além de questões técnicas, a própria participação dos indivíduos na pesquisa pode gerar dados ausentes. Pois, não é raro que participantes se recusem a fornecer informações que considerem sensíveis ou que alguns indivíduos não compareçam em determinados momentos da coleta (em estudos longitudinais), inviabilizando a obtenção de dados naquele período específico.

A melhor forma de lidar com os dados faltantes é prevenir o problema, planejando bem o estudo e coletando os dados de forma cuidadosa. Algumas estratégias para minimizar a quantidade de observações faltantes são sugeridas por (Scharfstein et al., 2012).

Por ser uma questão que é observada com muita frequência, Kang (2013) acrescenta que, uma das primeiras estratégias para lidar com a questão dos dados faltantes, é utilizar durante o processo de estimação, técnicas consideradas robustas, ou seja, métodos que permitem acomodar os problemas causados pela ausência de algumas informações nas variáveis. Por robustez nos referimos a métodos que nos permitem ter a confiança de que violações leves a moderadas das suposições produzirão pouco ou nenhum viés ou distorção nas conclusões tiradas sobre a população. No entanto, nem sempre é possível utilizar tais técnicas. Desta forma, uma série de alternativas para lidar com os dados faltantes são apresentadas na literatura.

Uma das principais abordagens, e talvez a mais acessível, é a omissão dos casos (ou indivíduos) com observações incompletas nas covariáveis. Essa abordagem é conhecida como análise

dos casos observados (ou casos disponíveis), ou simplesmente omissão dos casos ausentes. É importante salientar que, tal abordagem é a que está implementada por "*default*" em boa parte dos softwares estatísticos. Entretanto, apesar de ser a solução mais simplista para lidar com o problema dos dados faltantes, tal abordagem pode introduzir viés no processo de estimação. E a severidade (ou magnitude) desse viés pode aumentar dependendo do mecanismo que originou a perda dos dados, e/ou do tamanho de amostra (Hsu e Yu, 2019).

Para tratar adequadamente os dados faltantes, é essencial compreender o mecanismo que os gerou. Pois, tais mecanismos influenciam diretamente a forma como os dados ausentes podem ser tratados e imputados. Segundo Rubin e Little (2002), os dados faltantes podem ser classificados em três categorias principais: dados ausentes completamente ao acaso (*Missing Completely at Random*-MCAR), dados ausentes ao acaso (*Missing at Random*-MAR) e dados não ausentes ao acaso (*Missing Not at Random*-MNAR). Cada uma dessas categorias implica diferentes níveis de aleatoriedade e relação com outras variáveis do estudo, e sua correta identificação é fundamental para a escolha da técnica de imputação mais apropriada.

Neste trabalho, será utilizada a técnica Multiple Imputation by Chained Equations (MICE), por meio do mecanismo MAR, que pode ser vista como uma suposição razoável em pesquisas envolvendo dados de sobrevivência, em que o mecanismo de ausência dos dados não depende do tempo de censura, mas pode estar relacionado ao tempo de falha. Ademais, a técnica MICE é considerada uma das mais robustas para lidar com dados faltantes em análises complexas com as características supracitadas (White et al., 2011).

O presente estudo foi motivado pelo banco de dados de melanoma, inicialmente estudado por Cherobin et al. (2018). Em suma, melanoma é um tipo de câncer agressivo que surge quando os melanócitos, células produtoras de pigmento na pele, começam a proliferar descontroladamente. Estudos apontam que esse tipo de câncer é predominantemente observado em adultos brancos, cuja sua taxa de mortalidade é alta, principalmente quando diagnosticado em estágios avançados.

O diagnóstico precoce e a identificação de pacientes em grupos de alto risco para o desen-

volvimento de metástases estão entre as principais estratégias para reduzir a taxa de mortalidade associada a essa doença (Costa et al., 2024; Almeida et al., 2023b). Mesmo representando, no Brasil, somente 4% das neoplasias malignas da pele, esse tipo de câncer merece atenção especial devido à sua gravidade, pois, tem alta possibilidade de metástase (Costa et al., 2024). Além disso, em pesquisas realizadas com base nos dados do Departamento de Informática do SUS (DATASUS), percebe-se um aumento do número de casos ao longo dos anos no Brasil, com a região sul destacando-se por conter o maior número de casos por 100 mil habitantes (Rodrigues et al., 2024; Luz et al., 2023).

O banco de dados de melanoma é composto por observações coletadas entre 1995 e 2012 no Hospital das Clínicas da Universidade Federal de Minas Gerais, e no serviço privado de Oncologia Cirúrgica do Aparelho Digestivo (ONCAD) de Belo Horizonte. O mesmo, consiste em informações de 514 pacientes diagnosticados com melanoma cutâneo invasivo primário. O objetivo do estudo era de avaliar a influência de um conjunto de características epidemiológicas e histopatológicas no desenvolvimento da metástases em pacientes diagnosticados com melanoma.

Entretanto, entre os 514 pacientes que foram acompanhados durante os 17 anos, apenas 221 deles apresentaram informações completas nas covariáveis. O que significa que, mais da metade dos indivíduos ($\approx 57\%$) apresentaram informações incompletas (dados ausentes), em pelo menos uma das variáveis. Recentemente, vários trabalhos foram publicados, como fruto da modelagem do banco de dados de melanoma, nomeadamente, Cherobin et al. (2018), Almeida et al. (2018), Almeida et al. (2021), Almeida et al. (2022) e Almeida et al. (2023a).

É importante salientar que, nos trabalhos supracitados a modelagem considerou apenas os pacientes com informação completa nas covariáveis, o que pode, de alguma forma, nos levar a questionar do quão representativa foi a amostra utilizada para tal finalidade, bem como a questão do poder amostral, que em partes, pode ter influenciado as estimativas. Ademais, suspeita-se que, a exclusão dos casos ausentes possa ter ocasionado o problema de convergência do algoritmo de estimação, comumente conhecido na literatura, como *verossimilhança monótona*

(VM).

A VM é o problema de não existência de estimativas finitas para determinados coeficientes de regressão. Ou seja, quando a função de verosimilhança não é maximizada para determinados parâmetros, fazendo com que ela seja divergente (ou estritamente monótona para $\pm\infty$). Em linhas gerais, esse fenômeno pode ser observado em estudos envolvendo amostras pequenas com uma elevada proporção de censura.

A principal característica que pode induzir a ocorrência da VM é a presença de covariáveis categóricas binárias (com elevado desbalanço), quando todas as falhas estão associados a um único nível. Ou seja, segundo Andersen et al. (1985), Heinze e Schemper (2001), e Almeida et al. (2022), quanto maior for o número de covariáveis categóricas binárias na estrutura de regressão, maior é a chance da ocorrência desse fenômeno. No contexto envolvendo apenas covariáveis contínuas, as chances de ocorrência do problema reduzem significativamente, conforme apresentado em Klein e Moeschberger (2006).

No conjunto de dados em questão, o problema da VM é observado depois que os dados faltantes são descartados nas análises. Ou seja, para a covariável mitose, um dos fatores prognósticos de suma importância no modelo (Damato et al., 2011; Arce et al., 2014), mostrou ausência de metástase (evento em estudo) para os indivíduos que não desenvolveram a mitose. Ou seja, todas as falhas estão associadas a um único nível (*presença de mitose*), vide a Figura 1.1.

Esse resultado é um potencial indicador da presença da verossimilhança monótona no conjunto de dados. Maiores detalhes podem ser encontrados em (Almeida et al., 2018). O segundo desafio que pode ser observado na Figura 1.1, é a presença de uma parcela de sobreviventes de longa duração, ou seja, aqueles indivíduos que não são susceptíveis a ocorrência do evento em estudo. Esse segundo desafio foi tratado de forma exaustiva por Almeida et al. (2021) e Almeida et al. (2022), nos casos envolvendo pacientes com informação completa nas covariáveis.

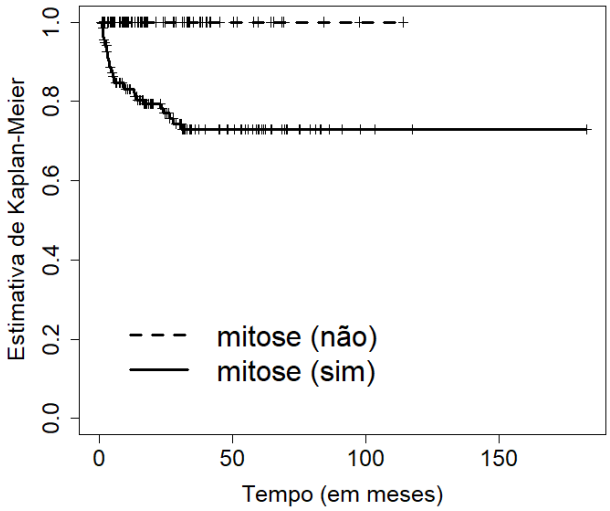

Figura 1.1: Curva de Kaplan-Meier para o fator mitose.

1.1 Contribuições

Considerando que a exclusão de unidades experimentais com observações incompletas nas covariáveis pode ter potencializado a ocorrência da VM, as principais contribuições do presente trabalho incluem: (i) aplicar os métodos de imputação de dados, como uma alternativa à exclusão dos valores faltantes; (ii) avaliar a eficácia e robustez dos métodos de imputação (capacidade em solucionar o problema da VM); (iii) desenvolver um estudo de simulação Monte Carlo considerando diferentes níveis de dados faltantes e diferentes percentagens de amostras com VM. Por fim, (iv) propor uma aplicação ao banco de dados de melanoma, para aferir o poderio dos métodos de imputação em atenuar (ou lidar) com o problema da VM.

1.2 Organização da dissertação

A presente dissertação está organizada da seguinte forma: o Capítulo 2 apresenta os conceitos básicos de análise de sobrevivência, o Capítulo 3 descreve as técnicas de imputação de dados, no Capítulo 4 apresentamos uma discussão geral sobre os estudo de simulação, e o Capítulo 5 apresenta os resultadas da aplicação dos conhecimentos obtidos a um banco de dados reais.

Capítulo 2

Análise de Sobrevivência

Neste capítulo serão apresentados os conceitos gerais de análise de sobrevivência. Na Seção 2.1, serão discutidos os conceitos básicos essenciais para a compreensão desta área de estudo e também será abordado o tempo de sobrevivência, analisando como esse fator é medido e interpretado em estudos de sobrevivência. Na Seção 2.2, será introduzido o modelo de Cox ou riscos proporcionais. Na Seção 2.3, será detalhado o método da máxima verossimilhança parcial. Na Seção 2.4, serão explorados tópicos em estimação inferencial de estatística em análise de sobrevivência.

2.1 Conceitos Básicos

A análise de sobrevivência envolve uma classe de técnicas estatísticas que se ocupam em estudar a variável tempo até que um determinado evento de interesse ocorra, também conhecido como tempo de falha. Apesar do nome, a aplicação da técnica de análise de sobrevivência não se restringe apenas a área de saúde. Esse evento de interesse pode estar associado às mais diversas áreas do conhecimento, dependendo da pesquisa, como economia, engenharia, saúde, educação, entre outras. Ou seja, a técnica possui ampla aplicação em diferentes áreas. Para citar alguns exemplos práticos que foram estudados: o tempo até a falência de bancos privados no Brasil

(Alves, 2009), o tempo até a falha de um equipamento (Papathanasiou et al., 2023), o tempo de recuperação de um paciente (Tolossa et al., 2021), o tempo até a morte de um paciente (Önal et al., 2020), o tempo até a evasão ou diplomação de estudantes em cursos de graduação (Lima Junior et al., 2012).

É importante destacar que os estudos em análise de sobrevivência acompanham o objeto de estudo (seja ele um indivíduo, entidade, etc.) ao longo do tempo, até que o evento de interesse ocorra, ou até que se torne inviável continuar o acompanhamento ou obter mais informações. Assim, a escala de medida da variável resposta pode variar conforme as demandas e necessidades da pesquisa, bem como a disponibilidade dos dados. Essa medida pode ser expressa em segundos, minutos, dias, meses, entre outras unidades de tempo. Como se trata de uma observação que se estende ao longo do tempo, diversos fatores podem interferir durante o estudo, como, por exemplo, a perda de acompanhamento de um paciente, o término do prazo da pesquisa, ou a retirada do objeto de estudo por razões alheias à vontade do pesquisador. Quando o acompanhamento é interrompido antes que o evento de interesse ocorra, esses casos são denominados dados censurados.

Em suma, censuras são observações parciais e incompletas de respostas e, segundo Hosmer et al. (2008) e Lawless (2003), podem ser classificadas em:

Censura à esquerda: ocorre quando a observação do indivíduo verifica-se após o acontecimento do evento de interesse. Nesse caso, apenas sabemos que o tempo de sobrevivência é inferior a um certo limite, mas não seu valor exato. Um exemplo disso seria um estudo em que os pesquisadores desejam determinar a idade em que as crianças de uma determinada comunidade começaram a ler. Pois, é comum observar que algumas crianças participantes da pesquisa já sabiam ler antes do início do estudo, caracterizando assim a censura à esquerda (Colosimo e Giolo, 2006).

Censura intervalar: Neste caso, sabe-se que o evento de interesse ocorreu em um determinado intervalo de tempo, mas o momento exato dentro desse intervalo não é conhecido. Por exemplo, em estudos onde os pacientes são acompanhados em visitas periódicas, pode-se ape-

nas inferir que o evento ocorreu entre uma visita e outra.

Censura à direita: ocorre quando o evento de interesse não foi observado até o final do período de acompanhamento, ou seja, sabe-se apenas que o tempo de sobrevivência é maior do que um certo limite. Esse é o tipo mais comum de censura, indicando que o evento ainda não ocorreu ou que o indivíduo foi perdido no acompanhamento. Retomando o exemplo da pesquisa sobre a idade em que as crianças começaram a ler (Colosimo e Giolo, 2006), algumas crianças ainda não tinham começado a ler até o final do período de acompanhamento do estudo. Nesse caso, sabe-se apenas que elas aprenderão a ler em algum momento após o término do estudo, caracterizando a censura à direita.

Para pesquisas que apresentam tanto censura à direita quanto à esquerda, como no caso ilustrado, os dados são denominados duplamente censurados (Turnbull, 1974).

A censura à direita é subdividida em outros três tipos, são eles:

Censura do tipo I: nesse caso, o estudo tem um período de tempo específico determinado pelo pesquisador. Transcorrido o tempo proposto para a pesquisa, independentemente se todos os objetos de estudo vivenciaram o evento de interesse ou não, o estudo é encerrado. Por isso, neste caso, o percentual de censuras é uma variável aleatória.

Censura do tipo II: neste caso, a quantidade de falhas é previamente definida. O pesquisador estabelece o número desejado de falhas que devem ocorrer, digamos k, e o estudo é encerrado assim que k falhas forem observadas. Assim, podemos dizer que o experimentador controla o número de falhas observadas, que será exatamente k, em uma amostra de tamanho n.

Censura aleatória: este é o caso mais geral e abarca os demais. Nele, os objetos de estudo não puderam ser acompanhados até o término do período proposto por motivos que estão além do interesse do estudo. Deste modo, por razões alheias à vontade do pesquisador, por exemplo, o paciente pode se ausentar da pesquisa, o equipamento deixa de ser acompanhado, informações sobre um elemento da amostra deixam de ser coletadas.

Para ilustrar os conceitos apresentados, observam-se as seguintes imagens:

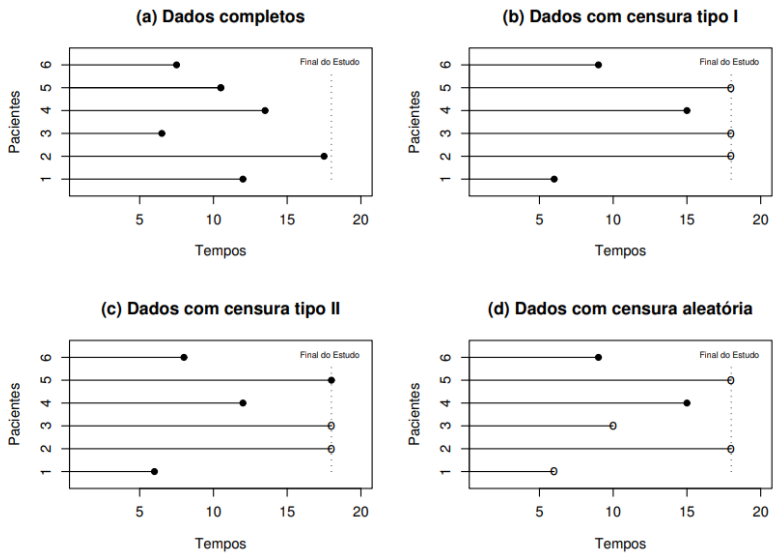

Figura 2.1: Apresentação de diversos mecanismos de censura, onde '•' indica a falha e '○' representa a censura. Fonte: Colosimo e Giolo (2006).

Deste modo, segundo Colosimo e Giolo (2006), é importante destacar que para cada indivíduo na amostra temos as seguintes variáveis: tempo até a ocorrência do evento de interesse (ou de falha) representada por t e a variável indicando se ocorreu a falha ou a censura representada por δ. Deste modo, tem-se:

$$\delta = \begin{cases} 1, & \text{se } t \text{ é tempo de falha} \\ 0, & \text{se } t \text{ é tempo de censura.} \end{cases} \tag{2.1}$$

Além do tempo de sobrevivência e da variável indicadora de censura, os dados também podem incluir covariáveis que representam a diversidade na amostra, como faixa etária, gênero, entre outras, bem como os procedimentos aos quais os participantes são submetidos. Muitas vezes, o interesse da análise recai sobre a relação entre o período de permanência e algumas dessas variáveis de interesse.

2.1.1 Tempo de Sobrevivência

Seja T uma variável aleatória não negativa e absolutamente contínua, que representa o tempo até a ocorrência do evento de interesse. É relevante observar que esta variável possui distribuição

de probabilidade e é geralmente descrita por meio da sua função de sobrevivência $S(t)$, função de risco (ou taxa de falha), $h(t)$ ou função densidade de probabilidade $f(t)$.

2.1.2 Função Densidade de Probabilidades

Seja T uma variável aleatória contínua não negativa, a função densidade de probabilidade, é dado pelo limite da probabilidade de ocorrência de um evento de interesse $[t, t + \Delta t)$ dividido pelo comprimento do intervalo (Nakano, 2017), isto é,

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

E função satisfaz as seguintes, para todo $t \geq 0$, as seguintes condições: $f(t) \geq 0$, e $\int_0^\infty f(t)\, dt = 1$.

2.1.3 Função de Sobrevivência

Conforme mencionado por Colosimo e Giolo (2006), a função de sobrevida $S(t)$ surge como uma das principais ferramentas probabilísticas empregadas na análise de sobrevivência. Ela representa a probabilidade de um indivíduo sobreviver mesmo depois de transcorrido um tempo t. Ou seja, a probabilidade de o indivíduo não experimentar o evento de interesse até o tempo t, dada por:

$$S(t) = P(T > t) = \int_t^\infty f(t)\, dt, \quad t \geq 0. \tag{2.3}$$

E neste caso, a função de sobrevivência é uma função não-crescente, absolutamente contínua (ou seja, $S(t)$ é uma função própria), tal que (Lawless, 2003): $\lim_{t \to 0} S(t) = 1$, e $\lim_{t \to \infty} S(t) = 0$.

2.1.4 Função de Taxa de Falha (ou Risco)

A função de taxa de falha (ou função de risco) denota o risco instantâneo de um indivíduo experimentar o evento de interesse entre os tempos t e $t + \Delta t$, dado que o indivíduo sobreviveu até t. Esta função é denotada por $h(t)$. Tratando-se de uma variável aleatória contínua, ela pode ser definida como a probabilidade condicional de um indivíduo falhar num intervalo de tempo, dado que sobreviveu até o tempo t dividido pelo comprimento do intervalo.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \quad t \geq 0. \tag{2.4}$$

Vale ressaltar que, $h(t)$ assume valores reais positivos e não limitada superiormente (Lawless, 2003).

É importante observar que de (2.4) pode-se obter a relação entre as principais funções consideradas em análise de sobrevivência. O primeiro passo para analisar dados de sobrevivência consiste em fazer uma síntese do conjunto de dados. Esse procedimento é realizado utilizando o estimador de Kaplan-Meier, que será apresentado na próxima seção.

2.1.5 Estimador de Kaplan-Meier

A funções apresentadas anteriormente podem ser obtidas por meio de dois métodos: paramétricos e não-paramétricos. O primeiro caso envolve a especificação de um modelo paramétrico para a distribuição dos tempos de sobrevivência, por exemplo: modelo exponencial, Weibull, log-normal, entre outros. Já o segundo caso estima a função de sobrevivência a partir dos dados observados, ou seja, não faz suposições sobre a forma da distribuição dos tempos de sobrevivência.

O estimador de Kaplan-Meier, também conhecido como estimador limite-produto, é um exemplo de uma técnica não paramétrica que fornece uma estimativa da função de sobrevivência sem assumir uma forma específica para a distribuição dos tempos de sobrevivência. Esta técnica estatística não-paramétrica foi proposta por Kaplan e Meier (1958), e $\hat{S}_{KM}(t)$ é um estimador

de máxima verossimilhança (EMV) para a função de sobrevivência.

$$\hat{S}_{KM}(t) = \prod_{j:t_{(j)} \leq t} \left[1 - \frac{d_j}{n_j}\right], \tag{2.5}$$

onde t_j é o j-ésimo tempo distinto e ordenado em que ocorre uma falha em um conjunto de dados de sobrevivência, n_j é o número de indivíduos que estão sob risco no o tempo t_j (inclusive) e d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo t_j, $j = 1, 2, ..., k$. É importante notar que o estimador de Kaplan-Meier é uma generalização da função de sobrevivência empírica, adaptada para lidar com dados censurados. No caso especial em que não há censura nos dados, a fórmula de Kaplan-Meier se reduz matematicamente à forma mais simples da função de sobrevivência empírica, que pode ser expressa como:

$$S(t) = \frac{\text{número de indivíduos com tempo de sobrevivência } > t}{\text{número total de indivíduos}}, \text{ para todo } t \geq 0. \tag{2.6}$$

Que pode ser definido como:

$$\hat{S}(t) = \prod_{t_j \leq t} \frac{n_j - d_j}{n_j}, \quad \text{para } t_{(j)} \leq t < t_{(j+1)}, \quad j = 1, 2, \ldots, k, \tag{2.7}$$

onde o último tempo de falha observado, quando $j = k$, $t_{(k+1)}$ seria o último tempo de falha.

Além das afirmações já feitas acima sobre o estimador de Kaplan-Meyer, é importante salientar que por deter a característica de ser um EMV, $\hat{S}_{KM}(t)$ é um estimador fracamente consistente e não viciado para amostras grandes. Sua consistência e normalidade assintótica foram mostradas por Breslow e Crowley (1974). Logo, $\hat{S}_{KM}(t) \overset{a}{\sim} \mathcal{N}(S(t), \text{Var}(\hat{S}_{KM}(t)))$, onde $\overset{a}{\sim}$ denota a convergência assintótica de $\hat{S}_{KM}(t)$.

2.2 Modelo de Regressão de Cox

Para as situações abordadas neste estudo, relacionadas a dados de sobrevivência com o objetivo de estudar o tempo até a ocorrência de um evento, o modelo de regressão de Cox tem sido amplamente utilizado como uma técnica viável para atender às demandas de análise dessas informações quando há covariáveis envolvidas. Este modelo, também conhecido como modelo de riscos proporcionais, é uma ferramenta poderosa para entender como diferentes fatores influenciam o tempo até a ocorrência de um evento específico.

O modelo de Cox incorpora características tanto de modelos paramétricos quanto de modelos não-paramétricos. Introduzido pelo estatístico britânico Cox (1972), esse modelo se destaca por sua flexibilidade e capacidade de fornecer informações significativas em estudos de sobrevivência.

O modelo de regressão de Cox é caracterizado pela sua função de risco, que é dada por:

$$h(t) = h_0(t) g\left(\mathbf{x}, \boldsymbol{\beta}\right), \tag{2.8}$$

onde

- $h(t|\mathbf{x})$ denota a função de risco no tempo t.
- $h_0(t)$ é a função de risco base, que depende apenas do tempo (ou a parte não-paramétrica do modelo).
- $g\left(\mathbf{x}, \boldsymbol{\beta}\right)$, denota a parte paramétrica do modelo de Cox, com $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)'$ representando o vetor dos coeficientes de regressão, e $\mathbf{x} = (x_1, \cdots, x_p)'$ é um vetor linha da matriz experimental (ou matriz de covariáveis), $\mathbf{X}_{n\times p}$. Vale ressaltar que $g\left(\mathbf{x}, \boldsymbol{\beta}\right)$ deve satisfazer a condição de não-negatividade

Comumente referenciada como função de ligação, a parte paramétrica representada na equação 2.8 por $g\left(\mathbf{x}, \boldsymbol{\beta}\right) = \exp\left(\mathbf{x}'\boldsymbol{\beta}\right) = \exp(\beta_1 x_1 + \cdots + \beta_p x_p)$, assume que as covariáveis têm um efeito multiplicativo no risco. Os coeficientes em $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)'$ são estimados a partir

dos dados e interpretados como logaritmos das razões de risco. Essa suposição de linearidade e multiplicatividade é o que confere ao modelo seu caráter paramétrico.

A parte não-paramétrica de (2.8), $h_0(t)$, não assume uma forma específica e é deixada completamente não especificada. Isso permite o modelo adaptar-se a diferentes formas de distribuição do tempo até o evento sem impor restrições rígidas, o que confere maior flexibilidade ao modelo. Ademais, com base na função de risco apresentada anteriormente, é possível obter as demais funções frequentemente consideradas em análise de sobrevivência.

O modelo de Cox também é conhecido como modelo de riscos proporcionais porque a razão entre as taxas de falha de dois indivíduos distintos é constante ao longo do tempo. Em outras palavras, para dois indivíduos i e j com vetores de covariáveis, $\mathbf{x}_i$ e $\mathbf{x}_j \in \mathbf{X}$, a razão de suas taxas de falha é dada por:

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{h_0(t)\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \exp\{(\mathbf{x}_i - \mathbf{x}_j)'\,\boldsymbol{\beta}\}, \tag{2.9}$$

o que demonstra que essa razão é independente do tempo. Por exemplo, se no início do estudo um indivíduo possui um risco de morte duas vezes maior que o risco de outro indivíduo, então essa proporção permanecerá constante durante todo o período de acompanhamento. Assim, a principal suposição para a aplicação do modelo de regressão de Cox é que as taxas de falha devem ser proporcionais.

2.2.1 Estimação de Parâmetros

Formulado o modelo de regressão de Cox, a etapa a seguir é a estimação dos parâmetros que caracterizam o modelo. Em linhas gerais, o método frequentemente considerado em análise de sobrevivência é o de *máxima verossimilhança*. Esse método pressupõe encontrar estimativas de $\boldsymbol{\beta}$, que maximizam a probabilidade de uma amostra particular ser observada.

Para tal, considere $\boldsymbol{\beta}$ o vetor de parâmetros desconhecidos (ou coeficientes de regressão). Se $\mathcal{D} = (t_i, \delta_i, \mathbf{x}_i)$, com $i \in \{1, 2, \cdots, n\}$ denota o conjunto de dados observados, onde t_i representa o tempo de sobrevivência do i-ésimo indivíduo, δ_i o status do indivíduo em t_i, podendo ser

falha ($\delta_i = 1$) ou censura ($\delta_i = 0$), e $\mathbf{x}_i$ o correspondente vetor de covariáveis. Podemos definir $f(t_i|\mathbf{x}_i, \boldsymbol{\beta})$ a função densidade de probabilidade, obtida a partir da equação (2.8), e $S(t_i|\mathbf{x}_i, \boldsymbol{\beta})$ a função de sobrevivência, respectivamente. Suas expressões correspondentes são dadas por:

$$\begin{aligned} S(t_i|\mathbf{x}_i, \boldsymbol{\beta}) &= S_0\left(t_i\right)^{g(\mathbf{x}_i, \boldsymbol{\beta})} \\ f(t_i|\mathbf{x}_i, \boldsymbol{\beta}) &= h_0(t_i) g\left(\mathbf{x}_i, \boldsymbol{\beta}\right) S_0\left(t_i\right)^{g(\mathbf{x}_i, \boldsymbol{\beta})}, \end{aligned}$$

onde $S_0(t_i) = \exp\left(-\int_0^{t_i} h_0(v_i) dv_i\right)$ denota a função de sobrevivência basal. Assim, sob a suposição de que o mecanismo de censura é não informativo, veja em Colosimo e Giolo (2006) e Klein e Moeschberger (2006) a função de verossimilhança completa (aquela que leva em conta a função taxa de falha basal) é dada por:

$$L(\boldsymbol{\beta}) \propto \prod_{i=1}^{n} [f(t_i|\mathbf{x}_i, \boldsymbol{\beta})]^{\delta_i} [S(t_i|\mathbf{x}_i, \boldsymbol{\beta})]^{1-\delta_i} = \prod_{i=1}^{n} [h(t_i|\mathbf{x}_i, \boldsymbol{\beta})]^{\delta_i} [S(t_i|\mathbf{x}_i, \boldsymbol{\beta})]. \tag{2.10}$$

Para uma abordagem mais detalhada, o leitor pode consultar as referências Colosimo e Giolo (2006), Klein e Moeschberger (2006), Lawless (2003), entre outras.

Para facilitar a derivação matemática e a otimização numérica na obtenção dos estimadores de máxima verossimilhança (EMV), utiliza-se o logaritmo da função de verossimilhança. Observe que não há perda de generalidade ao se maximizar $\ell\left(\boldsymbol{\beta}\right)$ em vez de $L\left(\boldsymbol{\beta}\right)$, visto que o logaritmo é uma transformação monotônica crescente. Portanto, maximizar a equação (2.10) equivale a maximizar o logaritmo da função de verossimilhança, dada por:

$$\begin{aligned} \ell\left(\boldsymbol{\beta}\right) = \log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{\delta_i \log h(t_i|\mathbf{x}_i, \boldsymbol{\beta}) + \log S(t_i|\mathbf{x}_i, \boldsymbol{\beta})\right\}, \\ &= \sum_{i=1}^{n} \left\{\delta_i \left[\log h_0(t_i) + \log g\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right] + g\left(\mathbf{x}_i, \boldsymbol{\beta}\right) \log S_0(t_i)\right\} + C \end{aligned} \tag{2.11}$$

Assim, o EMV $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ é obtido maximizando a log-verossimilhança em (2.11). Isto é,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \ell\left(\boldsymbol{\beta}\right). \tag{2.12}$$

Como ficou claro, a log-verossimilhança apresentada na equação 2.11 depende da função risco basal. Desta forma, a maximização da log-verossimilhança completa apresentada na Equação 2.11 pode ser conduzida de duas maneiras principais: (i) especificando uma forma paramétrica para a função de risco basal, $h_0(t_i; \boldsymbol{\theta})$, ou seja, assumindo uma estrutura funcional específica (como a de Weibull) que é governada por um vetor de parâmetros próprio ($\boldsymbol{\theta}$), permitindo que tanto $\boldsymbol{\beta}$ quanto $\boldsymbol{\theta}$ sejam estimados diretamente na maximização da verossimilhança; ou (ii) utilizando o método de perfilamento da log-verossimilhança. Este último consiste em tratar a função risco basal $h_0(t_i)$ como uma quantidade desconhecida que é eliminada do processo de otimização por meio de integração ou ajustes matemáticos apropriados (Johansen, 1983; Murphy e Vaart, 2000). Essa abordagem é útil quando se deseja evitar suposições explícitas sobre $h_0(t_i)$ e, ao mesmo tempo, obter estimativas confiáveis para os parâmetros de interesse $\boldsymbol{\beta}$. No entanto, um método prático para obter os EMV's no modelo de regressão de Cox é o método de máxima verossimilhança parcial, descrito a seguir.

2.3 Método da Máxima Verossimilhança Parcial

Para realizar inferência com base no modelo de regressão de Cox, é necessário estimar seus parâmetros. No modelo de Cox, o componente não-paramétrico, representado pela função de risco base, não precisa ser especificado. Isso inicialmente parece um obstáculo para a aplicação direta do método de máxima verossimilhança. No entanto, Cox (1972) propôs uma solução inovadora, conhecida como método de máxima verossimilhança parcial, que permite estimar os coeficientes das covariáveis sem a necessidade de especificar a função de risco base. Esse método denominado parcial porque, ao contrário da máxima verossimilhança completa, não requer a especificação da função de risco base $h_0(t)$. Esta característica confere ao modelo de Cox sua natureza semi-paramétrica.

A função de verossimilhança parcial é construída considerando apenas a ordem dos tempos de falha observados, desconsiderando a forma exata da função de risco base $h_0(t)$. Assumindo

que os tempos de sobrevivência estejam ordenados, $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$ a função de verossimilhança parcial é:

$$L_p(\boldsymbol{\beta}) = \prod_{k=1}^{d} \frac{\exp(\boldsymbol{x_k'}\boldsymbol{\beta})}{\sum_{j \in R(t_k)} \exp(\boldsymbol{x_j'}\boldsymbol{\beta})}, \tag{2.13}$$

onde $\boldsymbol{x_k}$ são as covariáveis do indivíduo que falhou no tempo $t_{(k)}$, $R(t_{(k)}) = \{i : t_i \geq t_{(k)}\}$ é o conjunto de indivíduos que estão em risco imediatamente antes de $t_{(k)}$, sendo o denominador da equação a soma sobre todos os indivíduos em risco, e d denota o número distinto de tempos de falha.

Nos casos em que há empates nos tempos de falha (isto é, múltiplos indivíduos apresentam o mesmo tempo de falha $t_{(k)}$), a verossimilhança parcial é ajustada utilizando métodos como: Correção de Breslow (Breslow e Crowley, 1974), Correção de Efron (Efron, 1977) ou Correção Exata (Kalbfleisch e Prentice, 2002).

É possível interpretar essa função de verossimilhança parcial 2.13 da seguinte forma: o numerador $\exp(\boldsymbol{x_k'}\boldsymbol{\beta})$ representa o risco relativo do indivíduo que falhou no tempo $t_{(k)}$, já o denominador: $\sum_{j \in R(t_{(k)})} \exp(\boldsymbol{x_j'}\boldsymbol{\beta})$ é a soma dos riscos relativos de todos os indivíduos que estavam em risco imediatamente antes de $t_{(k)}$. Deste modo, para encontrar os estimadores $\hat{\boldsymbol{\beta}}$ que maximizam a verossimilhança parcial, toma-se o logaritmo da função de verossimilhança parcial para simplificar os cálculos:

$$\ell_p(\boldsymbol{\beta}) = \sum_{k=1}^{d} \left(\boldsymbol{x_k'}\boldsymbol{\beta} - \log \left(\sum_{j \in R(t_k)} \exp(\boldsymbol{x_j'}\boldsymbol{\beta}) \right) \right). \tag{2.14}$$

A maximização desta função log-verossimilhança parcial pode ser realizada usando métodos numéricos, como o algoritmo de Newton-Raphson.

2.4 Inferência Estatística

As estimativas pontuais obtidas pelos métodos supracitados são partes da inferência estatística. Entretanto, além dos métodos de estimação pontual, podemos igualmente estar interessados em testar a significância dos coeficientes de regressão, $\beta_s \in \boldsymbol{\beta}$, com $s = 1, 2, \cdots, p$ bem como construir seus respectivos intervalos de confiança.

A influência do s-ésimo fator de risco no modelo é aferida por meio dos testes de hipóteses. Isto é, $H_0 : \beta_s = 0$ contra $H_1 : \beta_s \neq 0$. Note que a hipótese nula indica que a s-ésima covariável não é importante no modelo, enquanto que a H_1 contrária as conclusões da H_0. Sob certas condições de regularidade, descritas em Cox e Hinkley (1974) e Cordeiro (1992), é razoável assumir que $\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}_p\left(\boldsymbol{\beta}, Var\left(\hat{\boldsymbol{\beta}}\right)\right)$, com $Var\left(\hat{\boldsymbol{\beta}}\right) = I^{-1}\left(\hat{\boldsymbol{\beta}}\right)$ denotando a matriz de variância-covariância, temos que,

$$z = \frac{\hat{\beta}_s - \beta_s}{\hat{\sigma}_{ss}} \overset{a}{\sim} \mathcal{N}(0, 1), \tag{2.15}$$

onde a entrada genérica (s, s) de $I^{-1}(\hat{\boldsymbol{\beta}})$ é a variância estimada de $\hat{\beta}_s$, denotada por $\hat{\sigma}^2_{ss}$. O erro-padrão de $\hat{\beta}_s$ é, portanto, a raiz quadrada desta variância, $\hat{\sigma}_{ss} = \sqrt{\hat{\sigma}^2_{ss}}$. Assim, a H_0 será rejeitada a favor da H_1 se o $valor-p = 2P(Z > |z|)$ for menor que algum nível de significância α previamente especificado.

Ademais, a significância dos parâmetros pode ser aferida igualmente, por meio dos intervalos de confiança. Tais intervalos podem ser construídos usando a suposição da normalidade assintótica de $\hat{\boldsymbol{\beta}}$. Assim, um intervalo de confiança ao nível de $100(1-\alpha)$ de β_s, pode ser escrito da seguinte maneira:

$$IC_{(1-\alpha)}(\beta_s) = \left[\hat{\beta}_s - z_{\frac{\alpha}{2}}\hat{\sigma}_{ss}; \hat{\beta}_s + z_{\frac{\alpha}{2}}\hat{\sigma}_{ss}\right]. \tag{2.16}$$

Assim, a hipótese nula $H_0 : \beta_s = 0$ será rejeitada a favor da H_1, com um nível de significância α, se o valor 0 não estiver contido no intervalo de confiança de $100(1-\alpha)\%$ para β_s, ou seja, se $0 \notin IC_{(1-\alpha)}(\beta_s)$. É importante notar que esta inferência para $\boldsymbol{\beta}$ no modelo de Cox

é obtida através da verossimilhança parcial, uma técnica que tem a vantagem de não exigir a especificação de uma forma para a função de risco basal $h_0(t)$, como detalhado em Colosimo e Giolo (2006).

A inferência baseada na normalidade assintótica do estimador de máxima verossimilhança é a abordagem padrão para modelos de regressão. No entanto, a validade dessa abordagem depende de certas condições de regularidade que podem não ser atendidas na prática, especialmente em amostras pequenas ou com dados desbalanceados.

Um desses problemas é a ocorrência de verossimilhança monótona (VM), fenômeno no qual as estimativas de máxima verossimilhança não convergem para valores finitos, tornando a inferência padrão inválida. Para contornar essa limitação, foi destacado o método com a abordagem de máxima verossimilhança penalizada, originalmente proposta por Firth (1993) e popularizada para modelos de sobrevivência nos trabalhos de Heinze e Schemper (2001).

2.4.1 Correção de Firth

O método de Firth (Firth, 1993) foi originalmente proposto para reduzir o viés dos estimadores de máxima verossimilhança na classe dos modelos lineares generalizados. No entanto, sua adaptação para resolver o problema da verossimilhança monótona deve-se às pesquisas de (Heinze e Schemper, 2001; Heinze e Schemper, 2006) que mostrou o quão poderoso é o método para garantir a curvatura do perfil da função de verossimilhança, mesmo em situações envolvendo o fenômeno da VM. Em linhas gerais, o método é baseado na modificação da função escore (ou penalização da função de verossimilhança).

A correção de Firth pressupõem redefenir a função de verossimilhança parcial apresentada na equação (2.14), por meio da inclusão de um termo de penalidade, de tal forma que, a função de verossimilhança penalizada seja dada por:

$$\ell_p^*(\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta}) + 1/2 \log |\mathcal{I}(\boldsymbol{\beta})|, \tag{2.17}$$

com $|\mathcal{I}(\boldsymbol{\beta})|$ denotando o determinante da matriz de informação de Fisher, dada por: $\mathcal{I}(\boldsymbol{\beta}) = -\mathbb{E}\left(\frac{\partial \ell_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top \boldsymbol{\beta}}\right)$. Observe que, para distribuições da família exponencial com parametrização canônica, o termo de penalidade $1/2 \log |\mathcal{I}(\boldsymbol{\beta})|$ pode ser interpretado como sendo o logaritmo da priori invariante de Jeffreys (Jeffreys, 1946) largamente considerada em inferência Bayesiana. E portanto, a função de verossimilhança em 2.17) pode ser entendida como o logaritmo da distribuição à posteriori, e portanto, o estimador de máxima verossimilhança penalizado de $\hat{\boldsymbol{\beta}}^*$ de $\boldsymbol{\beta}$, pode ser visto como sendo a moda à posteriori. No entanto, apesar de toda similaridade entre os dois enfoques (Frequentista e Bayesiano), todo processo de estimação será baseado sob o enfoque clássico. É importante reforçar que o ajuste do modelo de regressão de Cox por meio da correção de Firth pode ser feito usando o pacote coxphf do R (Ploner e Heinze, 2015).

Com base na expressão apresentada em (2.17), a função de verossimilhança penalizada é dada por: $L_p^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) |\mathcal{I}(\boldsymbol{\beta})|^{1/2}$. Maiores detalhes podem ser encontrados em Heinze e Schemper (2001) e Firth (1993). Por fim, a distribuição assintótica de $\hat{\boldsymbol{\beta}}^*$ é normal centrada em $\boldsymbol{\beta}$ com matriz de variância-covariância dada por $\mathcal{I}^{-1}\left(\hat{\boldsymbol{\beta}}^*\right)$.

Capítulo 3

Imputação de Dados

A ausência de respostas dos participantes de uma pesquisa gera dados faltantes, o que representa um desafio significativo para a análise dos resultados. Na concepção de He (2010), diversos podem ser os motivos que levam à indisponibilidade de dados: o preenchimento inadequado de algum item por parte dos entrevistados, recusa em responder determinada questão e até a perda do indivíduo ao longo do estudo, no caso de estudos longitudinais (Nunes et al., 2010). Com isso, os objetivos da pesquisa podem ser comprometidos. Dependendo da proporção de dados faltantes, a pesquisa pode se tornar inviável ou, no mínimo, os resultados obtidos podem sofrer uma redução na eficiência das estimativas (Haukoos e Newgard, 2007). Além disso, a quantidade excessiva de dados faltantes pode impedir o uso de técnicas convencionais de análise que assumem um conjunto de dados completos. Ademais, a inclusão de indivíduos com informação incompleta pode introduzir vieses à pesquisa, especialmente se houver um perfil específico de quem não respondeu a determinados itens. Esses vieses são difíceis de identificar, pois os motivos subjacentes à ausência de respostas muitas vezes não são conhecidos ou explorados, o que agrava a complexidade da análise (Rubin, 1987).

O problema de dados faltantes é uma preocupação relevante em diversas áreas de estudo, incluindo ciências sociais, medicina e pesquisa de mercado (Schafer e Graham, 2002). A falta de dados pode reduzir a eficiência dos resultados obtidos, exigindo a aplicação de técnicas

apropriadas para mitigar seus efeitos. Para isso, é importante entender a ausência de dados de modo a encontrar as técnicas mais adequadas para contorná-los. O uso inadequado dessas técnicas pode gerar mais problemas do que soluções para a pesquisa, causando distorções em estimativas, erros padrão e testes de hipóteses (Rubin e Little, 2002).

É importante conhecer os conceitos que envolvem os mecanismos e os padrões de dados faltantes. Os mecanismos referem-se às diferentes formas e razões pelas quais os dados podem estar ausentes em um banco de dados. Os padrões, por sua vez, dizem respeito à maneira como os dados faltantes estão distribuídos no banco de dados. Compreender esses conceitos auxiliará na definição de como os dados deverão ser adequadamente tratados, indicando quais técnicas de imputação são as mais adequadas para cada caso.

3.1 Mecanismos de Dados Faltantes

Geralmente, os mecanismos de dados faltantes são classificados em três categorias principais (Rubin, 1987): dados ausentes completamente ao acaso (*Missing Completely at Random*-MCAR), dados ausentes ao acaso (*Missing at Random*-MAR) e dados não ausentes ao acaso (*Missing Not at Random*-MNAR).

3.1.1 Mecanismo MCAR

Dados ausentes completamente ao acaso (*Missing Completely at Random*-MCAR) ocorre quando o motivo da perda de dados não está relacionado a nenhuma das variáveis observadas ou não observadas no estudo. Isso implica que a probabilidade de ausência de um dado é a mesma para todos os casos, independentemente das respostas dadas (ou não dadas) pelos indivíduos pesquisados. Pode-se citar como exemplo o caso apresentado por Bergamo (2007) em uma pesquisa hipotética relacionada ao peso dos entrevistados. Se a ausência de dados não estiver relacionada ao peso da pessoa ou a nenhuma das variáveis coletadas, como idade ou sexo, então o mecanismo de dado faltante para este caso pode ser classificado como MCAR.

3.1.2 Mecanismo MAR

Dados ausentes ao acaso (*Missing at Random*-MAR) ocorre quando a ausência de dados em uma determinada variável pode ser explicada por outras variáveis observadas no banco de dados. Isso significa que a probabilidade de um dado estar ausente está relacionada a essas outras variáveis, mas não à própria variável faltante. Retornando ao caso do estudo sobre peso, exemplificado por Bergamo (2007), suponhamos que as participantes do sexo feminino com idade mais avançada não se sintam à vontade para responder à pergunta sobre o peso. Nesse caso, a ausência desses dados poderia ser explicada por outras variáveis presentes no banco de dados.

3.1.3 Mecanismo MNAR

Nos dados não ausentes ao acaso (*Missing Not at Random*-MNAR) a probabilidade de valores faltantes é afetada pelo próprio dado ausente. Se as pessoas com sobrepeso tendem a não responder à pesquisa, então a ausência do dado estaria relacionada à própria variável de interesse, neste caso, o peso (Bergamo, 2007).

3.1.4 Mecanismos de Dados Ignoráveis e Não-Ignoráveis

Os dados ausentes podem ser considerados ignoráveis nos casos em que o mecanismo que gerou essa ausência não precisa ser modelado explicitamente para garantir inferências estatísticas corretas (Buhi et al., 2008). Isso ocorre nos casos de mecanismos MCAR e MAR. No caso de MCAR, os dados faltantes ocorrem de maneira completamente aleatória e não estão relacionados às variáveis observadas ou não observadas, permitindo que a análise seja realizada sem introduzir viés. Para o mecanismo MAR, embora a ausência de dados esteja relacionada a variáveis observadas, ela não está relacionada aos próprios valores faltantes, o que também permite o uso de métodos estatísticos convencionais com ajustes apropriados para as variáveis observadas.

Por outro lado, no caso de MNAR, o mecanismo é classificado como não-ignorável. Isso

significa que a ausência de dados está diretamente relacionada aos próprios valores faltantes, e, portanto, há necessidade de modelar explicitamente o mecanismo que causou a ausência dos dados. Ignorar esse mecanismo pode resultar em vieses significativos e em inferências incorretas, tornando indispensável o desenvolvimento de modelos que incorporem esse aspecto na análise.

3.2 Padrões de Dados Faltantes

O padrão de dados faltantes diz respeito à estrutura ou à forma como os dados estão distribuídos ao longo de um banco de dados. Ou seja, em um conjunto de dados, geralmente as linhas representam os indivíduos que participam de uma pesquisa, enquanto as colunas correspondem às variáveis que caracterizam esses indivíduos. Com isso em mente, é possível visualizar os dados em formato matricial, o que permite identificar como os dados ausentes estão estruturados ao longo do banco de dados. Os padrões típicos de dados faltantes estão descritos e ilustrados a seguir (Enders, 2010).

Padrão univariado: em casos assim os dados faltantes são verificados em apenas uma das variáveis do banco de dados, ou seja, os dados faltantes são verificados em apenas uma das colunas.

Padrão monótono: ocorre quando, para uma dada ordenação das variáveis, se o valor de uma delas está ausente, os valores de todas as variáveis subsequentes também estão. Este padrão é comum em pesquisas longitudinais, como em estudos clínicos onde ocorre o abandono do acompanhamento por parte dos pacientes.

Padrão não-monotônico (geral): também conhecido como arbitrário, ou completamente casual, indica um comportamento aleatório ao ter distribuído, ao longo de todo o banco de dados, valores ausentes.

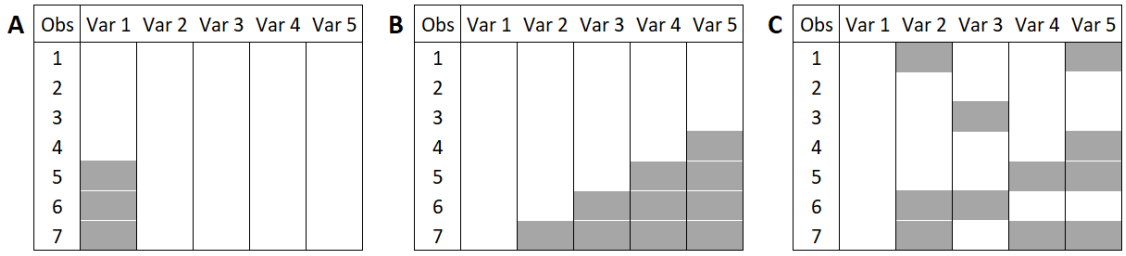

Figura 3.1: Padrões de ocorrência de dados faltantes. As células sombreadas representam dados faltantes. Temos em: (A) padrão univariado, (B) padrão monotônico, (C) padrão não monotônico (geral). Obs = observação; Var = variável. Fonte: adaptado de Haukoos e Newgard, 2007.

3.3 Técnicas para Tratar Dados Ausentes

É bastante recorrente em estudos e pesquisas o descarte de dados ausentes em bancos de dados. Por exemplo, essa é uma alternativa bastante utilizada na área de epidemiologia (Nunes et al., 2009). Mas essa não é a única alternativa que existe na literatura, além dela, também se utilizam as técnicas de imputação única. Dentre elas, parece existir um consenso de que não é recomendado a substituição de dados faltantes pela média ou mediana geral (Sinharay et al., 2001). Além dessas, também se apresenta como possibilidade a imputação múltipla.

3.3.1 Técnicas Não Sistemáticas

Na abordagem de análise de casos completos, apenas os indivíduos que não possuem dados faltantes em nenhuma das variáveis de interesse são incluídos na análise (Sinharay et al., 2001). Isso significa que qualquer indivíduo na pesquisa que tenha uma variável com informação faltante é excluído. Como possível consequência dessa prática, pode haver viés nas análises, já que os indivíduos excluídos podem ser diferentes daqueles que permaneceram. Além disso, vale destacar a redução do tamanho da amostra e o possível aumento do erro tipo II, pois o poder estatístico do teste é reduzido. Assim, pode-se pensar que as inferências obtidas por meio dessa análise podem estar comprometidas (Buhi et al., 2008). Contudo, apesar das diversas contraindicações, é possível que em casos nos quais os dados seguem um mecanismo MCAR,

alguns autores indiquem que essa técnica não causará viés nas análises (Haukoos e Newgard, 2007; He, 2010).

Outra alternativa utilizada por pesquisadores é a de trabalhar com os dados disponíveis. Nesse caso específico, são utilizadas todas as observações que possuem dados completos em determinada variável ou em um grupo específico de variáveis. Como consequência óbvia dessa prática, dependendo da análise ou das variáveis a serem estudadas, a quantidade de indivíduos observados pode variar de acordo com a disponibilidade de dados (Haukoos e Newgard, 2007; Sinharay et al., 2001). Embora essa possa parecer uma alternativa interessante, ela traz consigo os mesmos problemas da prática anterior. Portanto, em geral, pela praticidade da técnica, costuma-se utilizar apenas os dados completos na análise para simplificar o processo.

Uma outra possível abordagem é a de dados completos com ponderação (Haukoos e Newgard, 2007). Essa alternativa à análise de dados completos pondera os dados completos e incompletos de maneiras diferentes, com o objetivo de considerar potenciais vieses. Por um lado, ocorre uma redução do viés, o que é desejável; contudo, também há uma redução na precisão das estimativas, o que pode ser aceitável em bancos de dados grandes, onde a precisão das estimativas não é tão crucial (Haukoos e Newgard, 2007).

Uma possibilidade para realizar a ponderação dos dados é a técnica denominada Ponderação pela Probabilidade Inversa (ou Inverse Probability Weighting - IPW). Essa técnica funciona calculando a probabilidade de inclusão ou de ausência de dados para cada unidade na amostra e, em seguida, atribuindo pesos inversamente proporcionais a essas probabilidades. Ou seja, unidades com baixa probabilidade de inclusão ou alta probabilidade de dados faltantes recebem um peso maior, ajustando a análise para refletir melhor a população original ou para compensar o viés introduzido. No entanto, se houver dados faltantes em muitas variáveis e o padrão for não monotônico, pode ser inviável realizar uma análise utilizando essa técnica (Horton e Kleinman, 2007).

Existe também a técnica de imputação simples (ou imputação única), que é comumente utilizada na análise de dados devido à sua simplicidade de aplicação. Embora tenha suas li-

mitações, a facilidade e rapidez na implementação podem contribuir para torná-la uma técnica bastante acessível. A ideia é substituir valores faltantes por um único valor plausível que, em teoria, deve ser, ou pelo menos se aproximar, do valor real que ocuparia a célula (Haukoos e Newgard, 2007; He, 2010). Após a imputação, segue-se com as análises subsequentes como se o dado tivesse sido observado, ou seja, agora com um banco de dados completo. Vale ressaltar que existem diversas maneiras de gerar esse valor para substituir o valor ausente.

Uma das maneiras mais simples de realizar uma imputação única é utilizando a média (ou mediana). No entanto, é importante observar que, para realizar esse procedimento, os dados devem ser MCAR (Sinharay et al., 2001). Além disso, essa abordagem apresenta potenciais desvantagens para estudos e análises subsequentes. É crucial considerar que essa técnica pode reduzir a variabilidade dos dados, subestimando a verdadeira variância. Ademais, pode introduzir viés em estatísticas que se desviam da média, afetando a interpretação dos resultados. Portanto, recomenda-se utilizar essa técnica com cautela, especialmente em conjuntos de dados com poucos valores faltantes (Haukoos e Newgard, 2007; Buhi et al., 2008).

Uma alternativa mais sofisticada às técnicas de imputação única, como a substituição pela média, é o uso de modelos preditivos para estimar os valores ausentes (Sinharay et al., 2001). Essa abordagem emprega modelos estatísticos para estimar valores faltantes com base nas variáveis disponíveis no conjunto de dados. Modelos como regressão linear, árvores de decisão e redes neurais são exemplos comuns de técnicas utilizadas para esse fim. Os modelos preditivos procuram capturar relações mais detalhadas entre as variáveis.

3.3.2 Técnicas Baseadas em Verossimilhança

A técnica de máxima verossimilhança é amplamente utilizada na estimação de parâmetros de um modelo. Além disso, ela também pode ser aplicada para lidar com dados ausentes em um banco de dados (Allison, 2001). Maximizando a amostra observada, a técnica de máxima verossimilhança abarca propriedades como consistência, eficiência assintótica e normalidade assintótica, tudo isso sob diversas condições.

Por meio de técnicas que envolvem o método de máxima verossimilhança, é possível estimar os parâmetros de um banco de dados com dados ausentes, desde que o padrão de não resposta seja monotônico e o mecanismo de não resposta seja ignorável (Allison, 2001). Nesse caso, todas as informações disponíveis são utilizadas, os dados ausentes não precisam ser necessariamente imputados, e as estimativas podem ser mais eficientes, em alguns casos até mais eficientes do que ao utilizar a imputação múltipla (Graham et al., 2007).

Embora a máxima verossimilhança seja um princípio de estimação geral, sua aplicação em bancos de dados com valores ausentes representa um desafio computacional. A presença de uma grande quantidade de dados faltantes, em particular, pode dificultar a convergência de algoritmos iterativos (como o EM) que são utilizados para encontrar as estimativas de máxima verossimilhança(Haukoos e Newgard, 2007).

Para encontrar as estimativas de máxima verossimilhança em problemas com dados faltantes, uma abordagem computacional amplamente utilizada é o algoritmo EM (Expectation-Maximization). Essa técnica possui dois passos, sendo eles:

Expectation: Neste passo, calcula-se o valor esperado dos dados faltantes, condicionado aos dados observados e às estimativas atuais dos parâmetros. Quando se assume que os dados seguem uma distribuição Normal Multivariada, este passo se torna análogo à imputação por regressão, pois a esperança condicional dos dados faltantes é uma função linear das variáveis observadas.

Maximization: Partindo do banco de dados completamente preechido, utilizam-se os dados imputados no passo anterior para atualizar o vetor de médias e a matriz de covariâncias, que, por sua vez, serão usados para atualizar as estimativas dos dados faltantes. Em outras palavras, retorna-se ao passo anterior para gerar tais estimativas. Os dois passos são repetidos várias vezes até que se alcance a convergência (Allison, 2001; Nunes, 2007).

3.3.3 Técnicas de Imputação Múltipla

Como observado, a imputação simples oferece ferramentas básicas para preencher valores faltantes em um banco de dados. No entanto, é evidente que essas alternativas são frequentemente simplistas e, como consequência, podem acarretar problemas, conforme apresentado anteriormente.

Uma alternativa mais robusta de imputação de dados é a imputação múltipla, proposta por Rubin, 1987. Esse método possibilita (i) estimativas de parâmetros confiáveis (incluindo erros padrão); (ii) a estimação da informação faltante e seu impacto nas estimativas dos parâmetros; e (iii) pode ser aplicado em diferentes situações de dados faltantes (McKnight et al., 2007).

Nos métodos de imputação simples, os valores faltantes são preenchidos por um único valor calculado pelo método escolhido. Já na Imputação Múltipla, a ideia é calcular $m \geq 2$ valores para cada uma das lacunas no banco de dados. Dessa forma, são formadas m bases de dados completas, cujos resultados são agregados, e só então os valores faltantes podem ser preenchidos.

É importante salientar que vários métodos compõem essa família de técnicas chamada imputação múltipla, e saber selecionar aquele que mais se adequá ao banco de dados disponível é fundamental.

3.3.4 Passo a Passo da Imputação Múltipla

A Imputação Múltipla segue quatro passos principais: imputação, análise, agregação dos resultados e cálculo da informação faltante. O primeiro passo é a imputação. Um aspecto crucial nessa etapa é a quantidade de imputações, ou seja, o número de bancos de dados completos a serem gerados. Utiliza-se a letra m para representar essa quantidade. Portanto, m refere-se ao número de conjuntos de dados completos gerados a partir da imputação dos valores faltantes. Geralmente, m é escolhido como um número maior que 1 para permitir a variabilidade adequada entre os conjuntos imputados, e valores típicos variam entre 5 e 20 (Rubin, 1987).

Contudo, é importante salientar que a escolha de um m pequeno pode inflacionar o intervalo de confiança das estimativas e, consequentemente, reduzir o poder das análises (Graham et al., 2007). Ainda assim, exsitem referências na literatura mostrando que mesmo com um m modesto, como entre 3 e 5, é possível obter resultados suficientemente bons devido à eficácia da imputação múltipla, conforme (Molenberghs e Verbeke, 2006). Existem diversos métodos para gerar esses valores, e a escolha de um método adequado depende do tipo de dado disponível (como categórico nominal, categórico ordinal, numérico, entre outros).

O segundo passo é a análise, onde cada um dos m conjuntos completos é submetido a técnicas estatísticas apropriadas, resultando em m estimativas para cada parâmetro de interesse.

O terceiro passo é a agregação dos resultados. Nesta etapa, os múltiplos resultados obtidos no passo anterior são combinados em uma única estimativa final. A estimativa pontual final para cada parâmetro de interesse é calculada como a média das estimativas de cada banco de dados.

O quarto e último passo é o cálculo da incerteza total. Para obter a variância final e os intervalos de confiança, o método combina duas fontes de incerteza: aquela que já existiria se os dados fossem completos (a variância dentro de cada análise) e a incerteza adicional que surge por causa dos dados faltantes (a variância entre as diferentes análises). Essa combinação resulta em estimativas mais confiáveis, que levam em conta a ausência de dados.

A seguir as fórmulas necessárias para cada uma dessas etapas são apresentadas.

Fórmulas da Imputação Múltipla

Rubin (1987) descreveu algumas fórmulas simples para a fase de análise de vários conjuntos de estimativas de parâmetros e erros-padrão, com o objetivo de posteriormente combinar as análises dos m conjuntos de dados completos em um único conjunto de resultados. Neste contexto, o agrupamento das estimativas é dado por:

$$\overline{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j, \tag{3.1}$$

em que $\hat{Q}_j$ é a estimativa do j-ésimo parâmetro considerado, correspondente ao j-ésimo conjunto de dados imputados. Essas estimativas combinadas permitem obter uma única estimativa pontual, que consolida os resultados das análises realizadas nos m conjuntos imputados.
A combinação dos erros-padrão envolve duas fontes de variação:

- **A variância dentro das imputações** $(\hat{\sigma}_w^2)$: Definida como a média aritmética das m variâncias amostrais, descrita como:

$$\hat{\sigma}_w^2 = \frac{1}{m}\sum_{j=1}^{m}\hat{\sigma}_j^2, \tag{3.2}$$

  sendo $\hat{\sigma}_j^2$ a variância do j-ésimo conjunto de dados imputados.

- **A variação entre imputações** $(\hat{\sigma}_B^2)$: Que quantifica a variabilidade de uma estimativa em todas as m imputações, sendo simplesmente a variância do parâmetro estimado em todas as m imputações.

$$\hat{\sigma}_B^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Q}_j - \bar{Q})^2. \tag{3.3}$$

  Após o cálculo das estimativas combinadas $\bar{Q}$, $\hat{\sigma}_w^2$ e $\hat{\sigma}_B^2$, o próximo passo é a obtenção da variância combinada total, descrita por:

$$\hat{\sigma}_T^2 = \hat{\sigma}_w^2 + \left(1 + \frac{1}{m}\right)\hat{\sigma}_B^2. \tag{3.4}$$

Rubin (1987) também apresentou uma medida para determinar o incremento relativo da variância devido à presença das unidades ausentes:

$$r = \frac{(1+m)^{-1}\hat{\sigma}_B^2}{\hat{\sigma}_w^2}, \tag{3.5}$$

e uma taxa de unidades ausentes que se aproxima de

$$\lambda = \frac{r}{(1+r)}. \tag{3.6}$$

Quando há uma fração alta de dados faltantes, o número de conjuntos de dados a serem imputados deve ser determinado com maior cuidado. Nessas situações, um m maior reduz o impacto da variação entre imputações ($\hat{\sigma}_B^2$) na variância combinada total ($\hat{\sigma}_T^2$), resultando em estimativas mais estáveis. Por outro lado, estudos como os de White et al. (2011) destacam que o impacto de m na precisão das estimativas diminui conforme m aumenta. Mas frações muito altas de dados faltantes podem exigir valores de m acima do convencional. Esses autores sugerem que o aumento de m reduz o viés das estimativas, especialmente quando a fração de dados faltantes é alta. Como regra prática, para uma fração de dados ausentes superior a 50%, estudos como os de Graham et al. (2007) sugerem que m seja superior a 20.

Por fim, ressalta-se que, embora m maior aumente a precisão, ele também exige maior esforço computacional. Portanto, é essencial equilibrar a necessidade de precisão com as limitações práticas de recursos disponíveis.

3.4 Técnicas de Imputação Múltipla para Dados de Sobrevivência

Dados de sobrevivência podem ser tratados como um caso particular de dados ausentes, no qual o intervalo em que os valores não observados se encontram é conhecido. Quando o evento não ocorre, sabe-se que o tempo de sobrevivência excede o tempo de censura (quando se tem o mecanismo de censura à direita). Deste modo, o tempo de censura e o tempo de sobrevivência estão facilmente disponíveis, permitindo que a imputação seja direcionada para as covariáveis.

Deste modo, ao considerar a adoção de técnicas de imputação de dados faltantes em modelos de sobrevivência, o analista deve se atentar a vários aspectos: os pressupostos necessários para a imputação, a forma funcional do modelo, as variáveis incluídas, a quantidade de bancos de dados a serem gerados, a forma pela qual a imputação será realizada, além do desafio de integrar

adequadamente o tempo até a ocorrência do evento em estudo, bem como o indicador de falha nos modelos de imputação, de modo a manter a integridade da associação entre as covariáveis e o desfecho, evitando distorções nos resultados finais.

Para imputar dados faltantes em modelos de sobrevivência, é essencial incorporar corretamente o tempo de sobrevivência. Estudos indicam que a exclusão do tempo de sobrevivência pode comprometer a validade das estimativas (Van Buuren et al., 1999; White e Royston, 2009). Por exemplo, Ali et al. (2011) compara métodos de imputação que incluem o tempo de sobrevivência como preditor com aqueles que não o fazem. Entretanto, conclui-se que a exclusão do tempo de sobrevivência tende a introduzir vieses nos coeficientes dos modelos de regressão, favorecendo, assim, os modelos que consideram o tempo de sobrevivência, gerando resultados mais precisos.

Ao formular o modelo de imputação para uma variável incompleta X, considera-se sua relação com as variáveis completas $\mathbf{Z}$ e com o desfecho de sobrevivência, representado pelo vetor $(\mathbf{T}, \boldsymbol{\delta})$. Para prosseguir, assume-se que o mecanismo de censura é aleatório e não informativo, e que a distribuição do tempo de censura é independente de X condicionalmente a $\mathbf{Z}$.

A inclusão de $(\mathbf{T}, \boldsymbol{\delta})$ como preditores para X (sob esta suposição de independência) é justificada pois, mesmo com um mecanismo de censura independente, o tempo observado T (seja ele um tempo de falha ou de censura) fornece informações sobre o tempo de sobrevivência real, que por sua vez está associado a X. Portanto, utilizar $(\mathbf{T}, \boldsymbol{\delta})$ na imputação é crucial para se obter estimativas precisas e evitar viés.

Sob estas condições, a densidade conjunta de $(\mathbf{T}, \boldsymbol{\delta})$, segundo o modelo de Cox, é dada por:

$$f(\boldsymbol{t}, \boldsymbol{\delta} \mid x, z) \propto \left\{ h_0(t) e^{\beta x + \boldsymbol{\gamma}' z} \right\}^{\delta} \exp\left\{ -H_0(t) e^{\beta x + \boldsymbol{\gamma}' z} \right\}. \tag{4.1}$$

Para os parâmetros definidos pelo escalar β e o vetor $\boldsymbol{\gamma}$, o intuito é preencher os valores ausentes de X com base em sua distribuição condicional, levando em consideração as outras covariáveis e o desfecho de sobrevivência.

De acordo com White e Royston (2009), uma aproximação de $f(x|z,t,\delta)$ pode oferecer um modelo de imputação para X, seja ele categórico ou contínuo, ao incluir z, δ e $\hat{H}(t)$ como preditores lineares. Nesse contexto, $\hat{H}(t)$ é calculado pelo estimador de Nelson-Aalen. Deste modo, tempos em que os X serão imputados também são considerados na estimativa de $H(t)$. O motivo é que $H(t)$, obtido pelo estimador de Nelson-Aalen, é utilizado como um preditor no modelo de imputação para X. Assim, mesmo que alguns valores em X estejam ausentes, a imputação de seus valores faz uso de $H(t)$, que incorpora informações de todos os tempos disponíveis, sejam eles completos ou ausentes (através da imputação). Para a imputação de valores de X binário, adotamos um modelo logístico (se a variável faltante X não for binária, o método de imputação precisará ajustar o modelo condicional apropriado para o tipo de variável em questão):

$$\text{logit}\{P(X=1|z,t,\delta)\} = \eta_0 + \eta_1' z + \eta_2\delta + \eta_3\hat{H}(t), \tag{4.4}$$

onde $\text{logit}(p) = \log\frac{p}{1-p}$ representa as log-odds associadas à probabilidade p. Este modelo é ajustado utilizando o conjunto de dados atualmente imputado. As estimativas obtidas, juntamente com a matriz de covariância correspondente, podem ser empregadas como o vetor médio $\hat{\boldsymbol{\eta}} = (\hat{\eta}_0, \hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3)'$ e a matriz de covariância $\text{Var}(\hat{\boldsymbol{\eta}})$ de uma distribuição normal multivariada, possibilitando a geração de sorteios independentes $\boldsymbol{\eta^{(m)}}$ para $m = 1, \ldots, M$, onde M refere-se ao número de conjuntos imputados gerados no processo de imputação múltipla. Deste modo, K=m. Assim, para cada indivíduo com X ausente, um valor imputado $X^{(m)}$ pode ser extraído da distribuição de Bernoulli com a probabilidade de sucesso dada por:

$$P(X^{(m)} = 1 \mid z,t,\delta) = \frac{\exp\left(\eta_0^{(m)} + \eta_1^{(m)'} z + \eta_2^{(m)}\delta + \eta_3^{(m)}\hat{H}(t)\right)}{1 + \exp\left(\eta_0^{(m)} + \eta_1^{(m)'} z + \eta_2^{(m)}\delta_i + \eta_3^{(m)}\hat{H}(t_i)\right)}. \tag{3.7}$$

Nesta equação, $\eta_0, \eta_1, \eta_2, \eta_3$ são os parâmetros estimados do m-ésimo conjunto imputável e $\hat{H}(t)$ é a estimativa do risco acumulado, o que possibilita a imputação de valores para X com

base nos preditores z, δ e t.

Uma técnica utilizada para lidar com a imputação de dados faltantes que envolvem modelos de regressão de cox é o Multiple Imputation by Chained Equations (MICE), que tem demonstrado fornecer resultados satisfatórios e comparáveis a outras metodologias (Van Buuren et al., 1999; White e Royston, 2009). O processo de imputação pelo MICE segue um padrão semelhante ao de outras técnicas, envolvendo a seleção criteriosa das variáveis que irão compor o modelo e o procedimento geral de imputação.

Técnicas de imputação de dados tem suas limitações, ainda assim o MICE tem sido utilizado e testado por diversos estudos clínicos para casos de enfermidades como uma proposta plausível para tratar as lacunas deixadas pelos dados faltantes. Por exemplo, no artigo Grover e Gupta (2015) a técnica em questão é utilizada para imputar tanto covariáveis faltantes quanto desfechos censurados e, com isso, conseguiram estimativas melhores e erros padrão menores em comparação com outras abordagens, como a análise de casos completos e a exclusão de casos com covariáveis faltantes. Em Mbona et al. (2023), também observou-se que a técnica MICE gerou melhores estimativas, maior poder do modelo e menores erros padrão e um desempenho superior em comparação com o modelo de análise de casos completos. Além disso, também é importante observar a adaptabilidade dessa técnica quando aplicada a casos reais (Buuren e Groothuis-Oudshoorn, 2011).

Embora o MICE seja uma técnica robusta, é importante reconhecer suas limitações, especialmente em estudos com censuras. Modelos que não consideram adequadamente a censura ou a dependência temporal dos eventos podem introduzir vieses significativos. No entanto, pesquisas recentes demonstram que, quando aplicado corretamente e considerando as particularidades dos dados de sobrevivência, o MICE proporciona maior poder estatístico e estimativas mais confiáveis (Mbona et al., 2023).

3.5 Aplicação do MICE

Os primeiros relatos da aplicação do MICE surgiram em 1999, com o início de um procedimento chamado *regression switching* (Van Buuren et al., 1999). Nos anos seguintes, o MICE foi desenvolvido de forma mais consistente. O método é uma abordagem robusta para lidar com dados faltantes em conjuntos de dados complexos. Ele adota um processo iterativo para imputar valores ausentes, resultando em imputações mais precisas e confiáveis. A partir de alguns materiais (Azur et al., 2011; White et al., 2011), é possível resumir o procedimento geral do MICE em cinco etapas principais:

Etapa 1 (imputação inicial): O primeiro passo do MICE envolve a imputação inicial para preencher as lacunas do banco de dados. Esse processo usa uma imputação simples, gerando valores temporários para todos os dados faltantes. Esses valores servem como números provisórios iniciais.

Etapa 2 (imputação de cada variável , etapa iterativa): Em cada iteração, o MICE se foca em uma variável com valores ausentes de cada vez. Para a variável escolhida, seus valores imputados são temporariamente removidos, transformando-os novamente em dados faltantes. Essa variável se torna a variável dependente em um modelo de imputação.

Etapa 3 (construção de modelos univariados): Para a variável com dados faltantes, o MICE ajusta um modelo de imputação adequado, como regressão linear (para variáveis numéricas), regressão logística (para variáveis categóricas), etc. É importante reforçar que as variáveis do conjunto de dados (com ou sem valores imputados) são utilizadas como preditoras nesse modelo.

Etapa 4 (substituição dos valores faltantes): Os valores faltantes da variável-alvo são então substituídos pelos valores preditos pelo modelo. Uma vez imputada, essa variável é utilizada como preditora para imputar outras variáveis com dados ausentes nas iterações subsequentes.

Etapa 5 (ciclo de repetição): Esse processo de imputação continua iterativamente, passando de uma variável faltante para outra, até que as imputações se estabilizem após um certo número

de iterações (normalmente entre 5 e 20). Deste modo, o MICE baseia-se no princípio das Cadeias de Markov Monte Carlo (MCMC). Através desse princípio, são amostradas distribuições condicionais de cada variável, o que eventualmente permite a aproximação da distribuição conjunta dos dados.

3.5.1 Aplicação do MICE em R

Dentre a ampla gama de pacotes oferecidos pelo *software* R para imputação de dados, o MICE destaca-se como uma ferramenta popular para lidar com dados incompletos. Desde sua criação no início dos anos 2000, o MICE passou por várias atualizações até chegar à versão atual (geralmente referida como MICE 3.0+).

A função principal, `mice()`, é caracterizada por sua flexibilidade, sendo capaz de lidar com diferentes tipos de variáveis e oferecendo diversas opções de personalização. Isso permite que os usuários ajustem o processo de imputação conforme as necessidades específicas de suas análises. O MICE continua sendo uma escolha confiável para a imputação de dados ausentes em estudos epidemiológicos e em várias outras áreas de pesquisa, graças à sua combinação de flexibilidade e robustez, conforme destacado em White et al. (2011), Nguyen et al. (2017) e Linden e Adams (2016).

É importante destacar que o MICE exige a especificação do método de imputação para cada variável que possua dados ausentes. O método escolhido deve considerar a natureza da variável, levando em conta se ela é categórica nominal, ordinal, numérica discreta ou contínua. Abaixo, apresentamos algumas das técnicas disponíveis no pacote.

Dentre os diversos métodos apontados, vale ressaltar e explicar mais detalhadamente o método `polyreg`. Como descrito na Tabela 3.1, ele é utilizado para imputar variáveis categóricas com mais de duas categorias. O método emprega um modelo de regressão polinomial, especificamente um modelo de regressão multinomial, para lidar com dados categóricos (Brand, 1999).

O método `polyreg` ajusta um modelo multinomial para a variável categórica com dados

Tabela 3.1: Técnicas de imputação univariadas incorporadas no MICE.

Método	**Descrição**	**Tipo de Escala**
`cart`	Árvores de classificação e regressão	Qualquer
`lda`	Análise linear discriminante	Fator
`logreg`	Regressão logística	Fator, 2 níveis
`mean`	Imputação por média incondicional	Numérica
`norm`	Regressão linear bayesiana	Numérica
`norm.nob`	Regressão linear não bayesiana	Numérica
`norm.predict`	Regressão linear	Numérica
`pmm`	Matching de média preditiva	Numérica
`polr`	Modelo logístico ordenado	Ordenado
`polyreg`	Modelo logístico multinomial	Fator, >2 níveis
`sample`	Amostra aleatória dos dados observados	Qualquer

Fonte: RDocumentation e CRAN. Disponível em: *`https://cran.r-project.org/web/packages/mice/index.html`* *e* *`https://cran.r-project.org/web/packages/mice/mice.pdf`*.

faltantes. O modelo multinomial é uma extensão da regressão logística que permite modelar variáveis com mais de duas categorias sem uma ordem específica (Buuren e Oudshoorn, 2000). Após o ajuste do modelo, o método calcula as categorias previstas para as observações com dados faltantes com base nos valores das variáveis preditoras. Para garantir a variabilidade nas imputações e refletir a incerteza dos dados, o método adiciona um ruído apropriado às previsões das categorias.

A função `polyreg` utiliza a função `multinom()` do pacote `nnet` para ajustar o modelo multinomial. Para evitar viés devido a previsões imperfeitas, o algoritmo ajusta os dados de acordo com o método. Além disso, o método `polyreg` apresenta diversas vantagens. Por exemplo, uma vez que o modelo multinomial utilizado é capaz de capturar a complexidade das relações entre variáveis categóricas. Ademais, o método preserva a estrutura dos dados categóricos originais, o que é importante para análises subsequentes. A abordagem Bayesiana utilizada no `polyreg` ajuda a refletir a incerteza nas imputações, o que pode melhorar a robustez das análises estatísticas. No entanto, esse método possui igualmente algumas desvantagens. A principal delas é estar, em geral, associada a alto custo computacional, especialmente para variáveis

com muitas categorias. A criação de modelos multinomiais para variáveis com muitas categorias pode resultar em matrizes grandes e cálculos demorados. Isso pode limitar a escalabilidade do método e exigir mais recursos computacionais. Em comparação com outros métodos mais simples, como o `pmm`, o `polyreg` pode ser mais lento e mais exigente em termos de tempo computacional e memória, o que pode ser um fator a ser considerado dependendo do tamanho e da complexidade dos dados.

3.6 Especificação Condicional Totalmente Compatível com o Modelo Substantivo

O método de imputação MICE apresenta um desafio relevante: ele não leva em consideração o modelo de regressão de Cox no momento de gerar e calcular as imputações. Ainda assim, ele é amplamente citado em diversos artigos como um método eficiente para a imputação em modelos de sobrevivência. Deste modo, é possível aplicar a regressão de Cox em análises posteriores, mas isso não implica que o modelo seja essencial para o processo de imputação em si. Com essa limitação em mente, é pertinente apresentar uma alternativa que busca compatibilidade com o modelo em análise, que, neste caso, é o modelo de regressão de Cox.

O algoritmo *Substantive Model Compatible Fully Conditional Specification* (SMC-FCS), é um método avançado de imputação múltipla que garante a compatibilidade entre o processo de imputação e o modelo substantivo de análise. No entanto, uma limitação significativa deste método é a possibilidade de incompatibilidade entre o modelo de imputação e o modelo substantivo, para mais detalhes, veja Bartlett e Morris (2015)). Essa incompatibilidade ocorre quando não é possível definir um modelo conjunto que reproduza ambas as distribuições condicionais especificadas pelos modelos de imputação e substantivo. Como consequência, as estimativas do modelo substantivo podem ser viesadas, comprometendo a validade dos resultados. A compatibilidade é especialmente crucial em situações onde o modelo substantivo não é linear, como em modelos de sobrevivência exponenciais ou de Cox, frequentemente empregados para avaliar o tempo até a ocorrência de um evento.

3.6.1 Etapas do Algoritmo SMC-FCS

A principal inovação do SMC-FCS é sua capacidade de considerar explicitamente a relação entre as covariáveis parcialmente observadas e o resultado de interesse, garantindo que os valores imputados sejam consistentes com o modelo substantivo.

A primeira etapa do algoritmo é a especificação do modelo substantivo, ou seja, definir o modelo substantivo que representa a relação entre a resposta de interesse T e as covariáveis completas Z e parcialmente observadas X. No caso do modelo de Cox, o modelo substantivo assume a forma:

$$h(t|X, Z) = h_0(t) \exp(\psi_1 X_1 + \psi_2 X_2 + \cdots + \psi_p X_p + \psi_Z Z), \tag{3.8}$$

onde $h(t|X, Z)$ é a função de risco condicional, $h_0(t)$ é a função de risco basal, e ψ representa os coeficientes associados às covariáveis (Cox, 1972; Buuren, 2018).

Na segunda etapa, para cada covariável parcialmente observada X_j, um modelo de imputação é especificado. Este modelo deve capturar a relação de X_j com as outras covariáveis (X_{-j}) e as covariáveis completamente observadas (Z), bem como considerar a dependência de X_j no resultado T, conforme a seguinte expressão:

$$f(X_j|X_{-j}, Z, Y) \propto f(Y|X, Z, \phi) \cdot f(X_j|X_{-j}, Z, \phi_j), \tag{3.9}$$

em que:

- $f(T|X, Z, \phi)$ representa o modelo substantivo (Seaman et al., 2012),
- $f(X_j|X_{-j}, Z, \phi_j)$ é o modelo de imputação das covariáveis,
- ϕ e ϕ_j são vetores de parâmetros associados aos modelos substantivo e de imputação, respectivamente.

A especificação desse modelo garante que a imputação seja compatível com o modelo substantivo (Bartlett e Morris, 2015).

Na terceira etapa, todas as covariáveis parcialmente observadas X_j com valores faltantes são inicialmente preenchidas (imputadas) com estimativas preliminares. Essas estimativas podem ser obtidas de forma simples, como pela média, mediana ou imputação aleatória (Rubin, 1987).

Na quarta etapa, o algoritmo segue um processo iterativo, no qual cada covariável X_j com valores faltantes é atualizada de maneira condicional, enquanto as outras covariáveis (X_{-j}) e os parâmetros do modelo são mantidos fixos. Para cada X_j, os seguintes passos são realizados:

- **Cálculo da Densidade Condicional:** A densidade condicional $f(X_j|X_{-j}, Z, T)$ é avaliada com base na expressão descrita acima, considerando tanto o modelo substantivo quanto o modelo de imputação (White et al., 2011).

- **Extração de Amostras:** Valores para X_j são imputados amostrando-se da distribuição condicional proporcional a $f(X_j|X_{-j}, Z, T)$. Para variáveis categóricas, isso pode ser feito diretamente; para variáveis contínuas, técnicas como amostragem de Gibbs podem ser utilizadas (Van Buuren, 2018).

- **Atualização da Covariável:** Os valores imputados de X_j substituem os valores faltantes na base de dados.

O processo descrito acima é repetido iterativamente para todas as covariáveis parcialmente observadas $X_1, X_2, \ldots, X_p$. Cada covariável é atualizada em sequência, mantendo as imputações anteriores como fixas durante a imputação atual.

O processo iterativo continua até que algum critério de convergência seja alcançado. Esse critério pode ser baseado na estabilidade das imputações, na convergência dos parâmetros do modelo substantivo (Gelman e Rubin, 1992) ou num critério prático e comum para a convergência do algoritmo como a execução de um número pré-definido de iterações.

Assim, o processo descrito acima é repetido m vezes, gerando m conjuntos de dados imputados diferentes. Cada conjunto reflete a incerteza associada aos valores imputados (Rubin, 1987).

Após a imputação, o modelo substantivo (como o modelo de Cox) é ajustado separadamente a cada um dos conjuntos de dados imputados. As estimativas finais dos parâmetros de interesse são obtidas pela combinação das análises individuais, seguindo as Regras de Rubin (1987) já apresentadas neste trabalho na Subseção 3.3.4.

É importante reforçar que ao aplicar o método SMC-FCS no modelo de Cox, o algoritmo assegura que as imputações respeitem a estrutura do modelo de risco proporcional, preservando a relação entre as covariáveis e o tempo até o evento. Isso é particularmente importante para evitar viés nas estimativas dos coeficientes de risco e melhorar a precisão das inferências (Bartlett e Morris, 2015). Por exemplo, ao imputar uma covariável parcialmente observada X_j, o algoritmo considera explicitamente como essa covariável influencia a probabilidade condicional de sobrevivência (ou risco de evento) no modelo de Cox. Isso é realizado ao incorporar a função de risco proporcional $h(t|X, Z)$ no processo de imputação, garantindo que as imputações sejam substantivamente compatíveis.

É importante reforçar que, os métodos de imputação de dados não são técnicas *natas* para resolver o problema da verossimilhança monótona, em situações em que tal fenômeno pode ser observado. Ou seja, corrigir o problema dos dados perdidos não garante quebrar a estrutura da VM nos dados. Portanto, pode acontecer que, para determinados conjuntos de dados imputados a maximização da função de verossimilhança em (2.14) pode ser problemática em determinados regiões do espaço paramétrico (não existência de estimativas finitas para determinados coeficientes de regressão). Isto é, alguns conjuntos imputados podem preservar a estrutura de verossimilhança monótona (quando todas as falhas estiverem associadas a um único nível de uma covariável binária).

Com base no exposto anteriormente, um método que garante a obtenção de estimativas finitas para os coeficientes de regressão mesmo em situação envolvendo a VM e dados faltantes,

foi proposto por Firth, 1993, e mais tarde tornou-se popular devido aos trabalhos de Heinze e Schemper, 2001

3.6.2 Implementação em R

O pacote *smcfcs*, disponível no software estatístico R, implementa a imputação de valores ausentes em covariáveis utilizando a abordagem de Substantive Model Compatible Fully Conditional Specification (SMC-FCS), conforme proposta por Bartlett et al. (2015). Essa abordagem é especialmente adequada para situações em que o modelo substantivo inclui estruturas complexas, como efeitos não lineares, interações ou modelos de sobrevivência.

O pacote suporta imputação para diversos tipos de modelos, incluindo regressão linear ("lm"), regressão logística ("logistic"), regressão logística com viés reduzido ("brlogistic"), regressão de Poisson ("poisson"), Weibull ("weibull"), regressão de Cox para dados de tempo até o evento ("coxph") e modelos de Cox para dados de riscos concorrentes ("compet"). Para modelos de regressão de Cox, é necessário que o indicador de evento seja codificado como um valor inteiro, com "0"representando censura e "1"indicando a ocorrência do evento.

A função principal do pacote, *smcfcs*, retorna uma lista composta por dois elementos principais:

impDataset: uma lista contendo os conjuntos de dados imputados. Esses conjuntos podem ser utilizados para ajustar o modelo substantivo a cada banco de dados e, posteriormente, combinar os resultados utilizando as regras de Rubin por meio de pacotes como *mitools. smCoefIter*: uma matriz tridimensional que armazena os valores dos parâmetros do modelo substantivo obtidos ao final de cada iteração do algoritmo. A matriz é organizada de acordo com: o número de imputações, os parâmetros do modelo e o número de iterações. Nos casos em que o modelo substantivo seja de regressão linear, logística ou de Poisson, o *smcfcs* também realiza automaticamente a imputação de valores ausentes na variável resposta, caso existam. Contudo, mesmo nessa situação, é necessário que o usuário especifique no argumento correspondente ao método da variável resposta.

Adicionalmente, a estrutura de muitos dos argumentos da função *smcfcs* foi inspirada no renomado pacote *mice*, tornando o uso do *smcfcs* mais familiar para usuários já acostumados a trabalhar com imputação múltipla no R.

Capítulo 4

Estudo de Simulação

Neste capítulo, avaliamos o desempenho da técnica de imputação em termos de estimação por meio de estudos de simulações, no qual diferentes cenários são propostos. Todas as simulações e análises foram conduzidas no software R, versão 4.4.0, e foi utilizado processador Intel Core i7 e 16 GB de memória RAM.

Duas covariáveis foram consideradas na estrutura de regressão, tal que, para a i-ésima unidade experimental, segue que, $\mathbf{x}_i = (x_{1i}, x_{2i})^\top$. A primeira covariável, x_{1i}, é binária, fixada de tal forma que os primeiros $n_s = 10$ valores são iguais a 1 ($x_{1i} = 1$, denotando sucessos), e os demais $n_f = n - n_s$ iguais a 0 ($x_{1i} = 0$, denotando fracassos). A segunda covariável, x_{2i}, é contínua e foi gerada de uma distribuição normal padrão, com $i \in \{1, 2, \cdots, n\}$.

Os tempos de sobrevivência, T_i, foram gerados da distribuição Weibull (α, λ_i), com α denotando o parâmetro de forma, e $\lambda_i^* = \lambda \exp(\mathbf{x}_i^T \beta)$ denotando o parâmetro de escala, onde, $\lambda = \exp(\beta_0) = 1$. De igual forma, os tempos de censura, foram gerados de uma distribuição uniforma, tal que, $C_i \sim \mathcal{U}(0, \tau)$, onde τ é o parâmetro que controla a taxa de censura nos dados simulados.

O estudo de simulação envolveu dois cenários, respectivamente:

- **Cenário 1:** $\approx 15\%$ de amostras com VM: $\boldsymbol{\beta} = (-1,56;\ -0,22)^\top$.
- **Cenário 2:** $\approx 25\%$ de amostras com VM: $\boldsymbol{\beta} = (-2,00;\ 0,67)^\top$.

Nos dois cenários, os parâmetro α e τ foram fixados, tais que, $\alpha = 3,00$ e $\tau = 2,33$. De igual forma, a taxa de censura foi fixada em 50% para nos dois cenários.

As simulações de Monte Carlo foram realizadas com 1000 iterações para cada um dos seguintes tamanhos amostrais: $n \in \{50, 200, 500, 1000\}$. Para cada conjunto de dados simulado, as seguintes medidas de desempenho foram computadas: estimativas de Monte Carlo (eMC), que denota a média das estimativas dos coeficientes nas 1000 réplicas Monte Carlo, o viés relativo médio (VRM): $VRM = \frac{1}{1000}\sum_{q=1}^{1000}\left[\left(\hat{\beta}_q - \beta_q\right)/|\beta_q|\right]$, a raiz do erro quadrático médio ($rEQM$): $rEQM = \left[\frac{1}{1000}\sum_{q=1}^{1000}\left(\hat{\beta}_q - \beta_q\right)^2\right]^{1/2}$, o erro padrão Monte Carlo ($epMC$) e a probabilidade de cobertura empírica ($pCob$), para intervalos de 95% de confiança. Ademais, para avaliar a qualidade das imputações nos cenários com dados faltantes, foram utilizadas as seguintes métricas: a acurácia média de imputação para a covariável binária: x_1 (ACM), que segundo (Azur et al., 2011), essa medida representa a proporção de imputações corretas. Para a covariável contínua, x_2 temos as seguintes medidas: (i) erro absoluto médio (EAM); (ii) a raiz do erro quadrático médio ($rEQM$), que mede a magnitude média do erro de imputação; (iii) o desvio padrão da acurácia ($dpAC$); e o (iv) desvio-padrão da $rEQM$ (para avaliar a variabilidade do processo de imputação), conforme apresentado por (Rubin, 1987; White et al., 2011; Van Buuren, 2018).

Essas medidas permitem avaliar a precisão, o viés, a robustez e a qualidade das estimativas e imputações nos cenários de simulação estudados, os quais envolvem modelos de sobrevivência com dados censurados e faltantes. Para cada conjunto de dados gerado, todas as n unidades experimentais apresentaram informações completas nas covariáveis. Portanto, a contaminação dos dados com valores faltantes foi realizada para cada uma das variáveis consideradas na estrutura de regressão. O processo consistiu em selecionar $n\pi$ posições, de forma aleatória, independente, e sem reposição para cada uma das colunas do vetor de covariáveis $\mathbf{x}$, com $\pi \in \{0,1; 0,4\}$. Em seguida, as observações das covariáveis x_1 e x_2 presentes em cada uma das posições selecionadas foram substituídas por valores faltantes. Por fim, métodos de imputação

foram conduzidos, considerando 50 bancos de dados imputados. Posteriormente, a fórmula de Rubin apresentada na Seção (**??**) foi utilizada para agregar as estimativas obtidas nos conjuntos de dados completos.

Como os métodos de imputação não garantem eliminar o efeito da verossimilhança monótona em todos os conjuntos de dados imputados, a correção de Firth apresentada na seção (2.4.1), por meio do pacote coxphf do R. Os resultados de imputação apresentados nessa parte do trabalho foram baseados no método de imputação MICE.

Tabela 4.1: Resultados do modelo ajustado em conjuntos de dados sem valores faltantes em ambos os cenários. A correção de Firth foi considerada no ajuste. eMC : denota a estimativa de Monte Carlo; $epMC$: erro-padrão de Monte Carlo; VRM : vício relativo médio; $rEQM$: raiz quadrada do erro quadrático médio e $pCob$: probabilidade de cobertura.

Cenário 1: $\beta_1 = -1,56$ e $\beta_2 = -0,22$										
	eMC		VRM		rEQM		epMC		pCob	
n	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
50	-1,534	-0,235	0,017	-0,066	0,761	0,417	0,776	0,213	0,944	0,941
200	-1,431	-0,223	0,083	-0,014	0,745	0,282	0,700	0,095	0,928	0,933
500	-1,482	-0,221	0,050	-0,005	0,742	0,219	0,710	0,059	0,930	0,958
1000	-1,508	-0,220	0,033	-0,001	0,750	0,181	0,716	0,041	0,936	0,959
Cenário 2: $\beta_1 = -2,0$ e $\beta_2 = 0,67$										
	eMC		VRM		rEQM		epMC		pCob	
n	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
50	-1,950	0,690	0,025	0,030	0,826	0,453	0,961	0,249	0,940	0,959
200	-1,835	0,678	0,082	0,012	0,826	0,295	0,882	0,110	0,906	0,941
500	-1,847	0,673	0,077	0,004	0,807	0,233	0,866	0,068	0,913	0,952
1000	-1,855	0,672	0,072	0,003	0,800	0,191	0,869	0,047	0,913	0,961

Na Tabela (4.1) são apresentados os resultados do modelo ajustado em conjuntos de dados não contaminados (sem observações faltantes), para os dois cenários considerados no presente trabalho. Nela, é possível observar que o modelo apresenta um bom desempenho em termos de estimação, pois, as estimativas de Monte Carlo estão próximas dos valores reais. Ademais, as taxas de cobertura estão igualmente próximas do valor nominal para o coeficiente β_2 (em ambos os cenários) e, relativamente abaixo do valor nominal para o coeficiente β_1 (associado à covariável dicotômica), no segundo cenário. O vício relativo das estimativas dos dois coeficiente

é baixo, com destaque para as estimativas do coeficiente de regressão associado à covariável contínua que, em partes, mostrou não ser muito afetado pelo problema da VM, quando comparado com as estimativas do coeficiente associado à covariável binária e desbalanceada. Tal desbalanço associado ao fator x_1 pode ser esse aspecto que está influenciando os baixos níveis de cobertura. De igual forma, e como era de esperar, valores baixo e próximos foram obtidos para os erros-padrão das estimativas, quando comparados com os respectivos valores da raiz do erro quadrático médio, que varia de $0,742$ a $0,761$ (para $\hat{\beta}_1$) e $0,181$ a $0,417$ (para $\hat{\beta}_2$), no cenário 1, e $0,800$ a $0,826$ (para $\hat{\beta}_1$) e $0,191$ a $0,453$ (para $\hat{\beta}_2$), para o cenário 2. Para ambos os coeficientes de regressão, sua precisão tende a diminuir quando o tamanho de amostra aumenta.

Tabela 4.2: Resultados do cenário 1 (conjunto de dados contaminados com diferentes proporções de valores faltantes). Sendo eMC : a estimativa de Monte Carlo; $epMC$: erro-padrão de Monte Carlo; VRM : vício relativo médio; $rEQM$: raiz quadrada do erro quadrático médio e $pCob$: probabilidade de cobertura.

Caso 1: 10% de valores faltantes (Cenário 1: $\beta_1 = -1,56$ e $\beta_2 = -0,22$)										
	eMC		VRM		rEQM		epMC		pCob	
n	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
50	-1,474	-0,223	0,055	-0,015	0,771	0,223	0,809	0,230	0,935	0,971
200	-1,422	-0,214	0,088	0,025	0,714	0,096	0,741	0,101	0,920	0,965
500	-1,454	-0,221	0,068	-0,005	0,727	0,061	0,751	0,062	0,926	0,957
1000	-1,443	-0,220	0,075	0,001	0,721	0,041	0,743	0,043	0,930	0,961
Caso 2: 40% de valores faltantes (Cenário 1: $\beta_1 = -1,56$ e $\beta_2 = -0,22$)										
	eMC		VRM		rEQM		epMC		pCob	
n	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
50	-1,419	-0,195	0,090	0,113	0,832	0,251	0,985	0,274	0,941	0,980
200	-1,334	-0,209	0,145	0,049	0,765	0,121	0,899	0,122	0,918	0,951
500	-1,299	-0,213	0,168	0,030	0,744	0,073	0,872	0,075	0,902	0,950
1000	-1,305	-0,218	0,163	0,009	0,761	0,052	0,884	0,053	0,916	0,949

A Tabela 4.2 apresenta os resultados das principais medidas de desempenho avaliadas no presente trabalho. O modelo de regressão de Cox com verossimilhança monótona e dados faltantes foi ajustado para o cenário 1. Tal como no caso anterior (sem dados faltantes), valores baixos para o vício relativo das estimativas foram observados, para o Caso 1, no qual 10% das observações são faltantes (para cada um dos 1000 conjuntos de dados gerados). Medidas de

desempenho como rEQM, epMC e pCob apresentaram valores similares, quando comparado com aqueles reportados na Tabela (4.1). Já quando 40% dos dados são contaminados (40% de obervações faltantes), o desempenho das estimativas associadas à covariável binária reduziu significativamente, com VRM variando na faixa dos $0,090$ (quando $n = 50$) e $0,163$ (quando $n = 1000$). O epMC e rEQM apresentaram uma relativa diferença entre as duas medidas, como consequência do aumento do viés relativo. Por outro lado, a taxa de cobertura foi relativamente baixa, tendo estabilizado em torno dos 91%, aproximadamente. Uma vez mais, para o cenário em que a taxa de ocorrência de VM é relativamente baixa ($\approx$ 15% de amostras com VM), os coeficientes associados a covariável contínua apresentou um bom desempenho nos dois casos de contaminação de dados. As eMC está apresenta valores próximos dos reais (viés baixo), principalmente quando o n aumenta. Os valores da pCob estão em torno do valor nominal, e as estimativas de rEQM e epMC são similares, variando entre $0,043$ a $0,230$ (Caso 1), e $0,053$ a $0,274$ (Caso 2). Tais quantidades decrescem quando o tamanho de amostra aumenta.

Tabela 4.3: Métricas de qualidade da imputação MICE (Caso 1: 10% e Caso 2: 40% de valores faltantes), para o cenário 1. Sendo, ACM: a acurácia média, EAM : o erro absoluto médio, $dpAC$: o desvio padrão da acurácia, e $dp.rEQM$: o desvio padrão do erro quadrático médio.

Caso 1: 10% de valores faltantes					
	x_1		x_2		
n	ACM	dpAC	EAM	rEQM	dp.rEQM
50	0,694	0,105	1,156	1,377	0,235
200	0,896	0,038	1,124	1,392	0,120
500	0,955	0,017	1,117	1,393	0,075
1000	0,977	0,008	1,116	1,395	0,056
Caso 2: 40% de valores faltantes					
	x_1		x_2		
n	ACM	dpAC	EAM	rEQM	dp.rEQM
50	0,685	0,047	1,196	1,480	0,161
200	0,889	0,011	1,132	1,414	0,073
500	0,952	0,004	1,120	1,403	0,047
1000	0,975	0,002	1,116	1,398	0,033

Finalmente, o desempenho do modelo de regressão é corroborado pelas métricas de quali-

dade da imputação (Tabelas 4.3). Nela, observa-se uma melhora na qualidade do processo de imputação com o aumento do tamanho de amostra, verificada tanto no aumento da ACM para a variável binária quanto na redução do EAM e rEQM para a variável contínua, para os dois níveis da taxa de valores faltantes nos dados.

Tabela 4.4: Resultados do cenário 2 (conjunto de dados contaminados com diferentes proporções de valores faltantes). Sendo eMC : a estimativa de Monte Carlo; $epMC$: erro-padrão de Monte Carlo; VRM : vício relativo médio; $rEQM$: raiz quadrada do erro quadrático médio e $pCob$: probabilidade de cobertura.

Caso 1: 10% de valores faltantes (Cenário 2: $\beta_1 = -2,0$ e $\beta_2 = 0,67$)										
	eMC		VRM		rEQM		epMC		pCob	
n	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
50	-1,779	0,653	0,110	-0,026	0,818	0,261	0,985	0,271	0,927	0,967
200	-1,756	0,663	0,122	-0,011	0,773	0,112	0,915	0,117	0,914	0,954
500	-1,767	0,658	0,117	-0,018	0,774	0,070	0,912	0,072	0,918	0,959
1000	-1,753	0,657	0,124	-0,019	0,806	0,049	0,905	0,050	0,892	0,947
Caso 2: 40% de valores faltantes (Cenário 2: $\beta_1 = -2,0$ e $\beta_2 = 0,67$)										
	eMC		VRM		rEQM		epMC		pCob	
n	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
50	-1,576	0,564	0,212	-0,158	0,891	0,293	1,144	0,317	0,913	0,967
200	-1,438	0,609	0,281	-0,091	0,906	0,137	1,032	0,139	0,869	0,947
500	-1,435	0,621	0,283	-0,073	0,878	0,089	1,018	0,085	0,871	0,938
1000	-1,388	0,626	0,306	-0,066	0,925	0,070	1,001	0,060	0,854	0,903

Os resultados do ajuste do modelo para o Cenário 2 são apresentados na Tabela 4.4. Em linhas gerais, o desempenho das medidas avaliadas no presente trabalho piorou com o aumento da porcentagem de amostras com VM. Pois, o viés relativo estabilizou na faixa dos 12%, aproximadamente (Caso 1), e 27% (Caso 2). Ou seja, uma diferença considerável entre as eMC e o valores real de β_1 foi observada. Tal diferença acabou teve igualmente um reflexo nas estimativas de epMC e rEQM, em ambos os casos.

A taxa de cobertura variou de 0,93 (para $n = 50$) e 0,89 (para $n = 1000$), Caso 1 e 0,91 ($n = 50$) e 0,85 ($n = 1000$), para o Caso 2. Ou seja, com ou aumento do tamanho de amostra, a probabilidade de cobertura reduziu. Esse comportamento das estimativas de β_1 justifica-se pelo fato de que, como a covariável binária foi fixada em $n_s = 10$ (os primeiros 10 valores corres-

pondem o sucesso), e os demais $n_f = n - 10$ (os demais valores fracasso), portanto, quando o tamanho de amostra aumenta, o grau de desbalanço nos dados também aumenta. A título de exemplo, quando $n = 50$ temos 20% de sucessos em x_1, ao passo que, quando $n = 1000$, a percentagem de sucessos cai para 1%. Ademais, além do desbalanço supracitado, a contaminação dos dados gerados para os valores faltantes faz com que a qualidade das estimativas deteriore, pois, para esse mesmo Cenário, o modelo baseado em dados não contaminados (resultados da Tabela 4.1), são relativamente melhores do que aqueles reportados na Tabela 4.4.

Tal como os resultados do cenário anterior, a qualidade das medidas de desempenho das estimativas de β_2 são igualmente notáveis nesse segundo cenário. Por exemplo, quando 10% das observações são contaminadas (Caso 1), o vício relativo diminui com o aumento do tamanho de amostra, tendo seus valores variando de -0,011 a -0,026. O epMC e rEQM apresentam valores similares (como consequência dos baixos valores para o VRM), e a taxa de cobertura ronda em torno do valor nominal, isto é, de 0,967 (para $n = 50$) para 0,947 (para $n = 1000$).

Tabela 4.5: Métricas de qualidade da imputação MICE (Caso 1: 10% e Caso 2: 40% de valores faltantes), para o cenário 2. Sendo, ACM: a acurácia média, EAM : o erro absoluto médio, $dpAC$: o desvio padrão da acurácia, e $dp.rEQM$: o desvio padrão do erro quadrático médio.

Caso 1: 10% de valores faltantes					
	x_1		x_2		
n	ACM	dpAC	EAM	rEQM	dp.rMSE
50	0,688	0,104	1,092	1,303	0,231
200	0,896	0,037	1,038	1,285	0,112
500	0,955	0,017	1,030	1,285	0,070
1000	0,977	0,008	1,029	1,287	0,051
Caso 2: 40% de valores faltantes					
	x_1		x_2		
n	ACM	dpAC	EAM	rEQM	dp.rMSE
50	0,694	0,105	1,156	1,377	0,235
200	0,896	0,038	1,124	1,392	0,120
500	0,955	0,017	1,117	1,393	0,075
1000	0,977	0,008	1,116	1,395	0,056

Um aspecto que chamou atenção, e que deve ser referenciado aqui, está relacionado com a

qualidade das estimativas de β_2 para o segundo caso. Ou seja, em termos do viés relativo, ficou claro que tal quantidade diminui com o aumento do tamanho de amostra. Porém, a magnitude dessa redução é menor que aquela observada para o Caso 1. As probabilidades de cobertura estão em torno do valor nominal quando n é pequeno, e decresce, estabilizando em $0,903$ quando $n = 1000$. Por enquanto, não temos nenhuma explicação do porquê isso está acontecendo, pois, a força da VM não é refletida com tamanha magnitude nos coeficientes de regressão associados a covariáveis contínuas. Mas, suspeitamos que isso esteja acontecendo devido ao impacto do método de imputação considerado no presente trabalho. No que tange ás métricas de imputação, apresentaram estimativas similares com aquelas apresentadas no cenário 1.

Capítulo 5

Aplicação a dados reais

5.1 Dados de melanoma cutânea

O banco de dados analisado no presente estudo refere-se a informações de 514 pacientes, coletadas entre 1995 e 2012, no Hospital das Clínicas da Universidade Federal de Minas Gerais e no serviço privado de Oncologia Cirúrgica do Aparelho Digestivo (ONCAD) de Belo Horizonte. Diversas variáveis foram coletadas com o objetivo de entender sua possível influência no tempo até o desenvolvimento de metástases em pacientes diagnosticados com melanoma primário.

O banco de dados é composto por duas variáveis principais: uma variável numérica para o tempo de acompanhamento dos pacientes, medido em semanas, e uma variável binária para o desfecho da análise de sobrevivência. Esta, indica o status da presença de metástase (0: ausência, 1 presença). Além disso, o conjunto de dados inclui outras oito variáveis explicativas categóricas que foram utilizadas na modelagem.

As variáveis analisadas no estudo foram as seguintes: sexo biológico do paciente (1: masculino, 2: feminino); grupo etário (1: 18-40 anos, 2: 40-60 anos, 3: $>$ 60 anos); a localização primária do melanoma (1: cabeça, pescoço e tronco; 2: membros superiores e inferiores; 3: acral); e o tipo histológico (1: superficial expansivo e lentigo maligna melanoma, 3: nodular, 4: acral); profundidade de invasão do tumor avaliada pela categoria de Breslow (1: $\leq$1mm, 2: 1,01

a 4mm, 3: > 4mm); por fim, foram consideradas a presença ou ausência de ulceração (1: Não, 2: Sim) e de mitose (1: Não, 2: Sim). É possível visualizar como cada uma dessas variáveis estão distribuídos em cada uma de suas categorias conforme apresentado na Tabela 5.1.

Tabela 5.1: Distribuição das variáveis categóricas e dados faltantes no conjunto de dados.

Variável	Categoria	Frequência	Percentual (%)
Breslow	$\leq$1mm	236	45,91
	1,01 a 4mm	155	30,16
	$>$4mm	45	8,75
	Faltantes	78	15,18
Sexo	Masculino	218	42,41
	Feminino	295	57,39
	Faltantes	1	0,19
Grupo Etário	18-40 anos	124	24,12
	40-60 anos	195	37,94
	$>$60 anos	185	35,99
	Faltantes	10	1,95
Localização	Cabeça, pescoço e tronco	274	53,31
	Membros sup. e inf.	145	28,21
	Acral	85	16,54
	Faltantes	10	1,95
Metástase	Ausente	136	26,46
	Presente	378	73,54
Mitose	Não	87	16,93
	Sim	205	39,88
	Faltantes	222	43,19
Tipo Histológico	Superficial e Lentigo Maligno	240	46,69
	Nodular	74	14,40
	Acral	44	8,56
	Faltantes	156	30,35
Ulceração	Não	201	39,11
	Sim	70	13,62
	Faltantes	243	47,28

Algumas observações importantes devem ser destacadas com relação a essas variáveis. No caso da variável Sexo, foi identificado um registro com o valor `3`, o que não corresponde às categorias esperadas (masculino e feminino). Esse valor foi considerado um erro de codificação e, portanto, tratado como dado ausente (`NA`) durante o pré-processamento. Situação semelhante

ocorreu com a variável tipo histológico, que apresentou apenas uma observação com o valor `2`, o qual não se enquadra nas categorias relevantes do estudo. Por esse motivo, essa ocorrência também foi recodificada como `NA`.

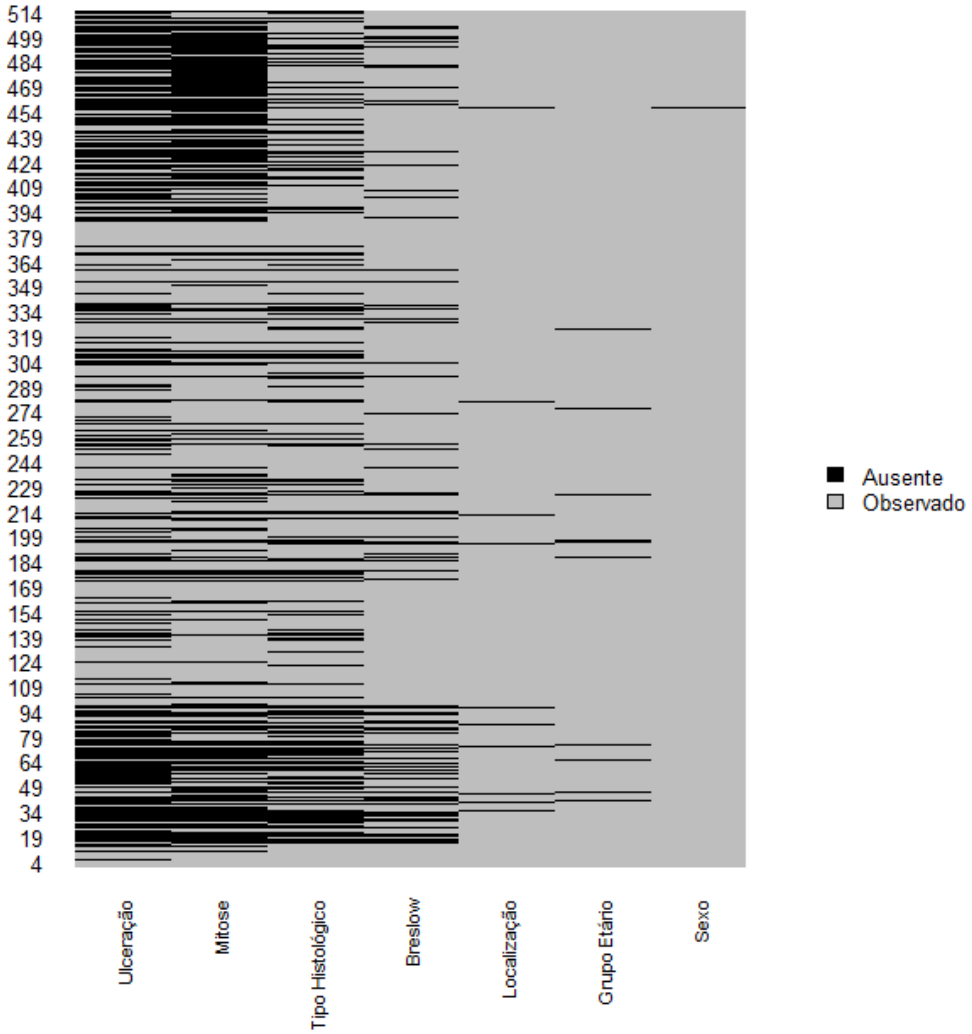

Figura 5.1: Mapa de dados faltantes, destacando as células com dados faltantes e observados. Fonte: Elaboração própria

Além disso, com o objetivo de destacar a ocorrência de dados ausentes no conjunto analisado, os resultados da Tabela 5.1, mostram claramente que fatores como Tipo histológico, Mitose e Ulceração contém uma taxa significativa de observações faltantes. Outra forma de

visualizar os dados ausentes é por meio de um mapa de dados conforme apresentado na Figura 5.1, o qual destaca graficamente as células observadas e faltantes no conjunto, facilitando a identificação de possíveis padrões na ausência dos dados e também, permitindo avaliar se os valores ausentes ocorrem de forma aleatória ou se seguem algum comportamento sistemático. Além disso, tanto com a tabela quanto com a figura é possível perceber os altos percentuais de valores ausentes para as variáveis desse banco de dados.

Além da análise dos valores ausentes, a curva de Kaplan-Meier apresentada na Figura 1.1 foi obtida considerando apenas os casos completos para a variável mitose. O comportamento observado nesse caso é indicativo da presença de verossimilhança monótona no conjunto de dados, conforme discutido por Almeida et al. (2018).

5.1.1 Imputação de dados no banco melanoma

Foram utilizadas duas estratégias distintas de imputação: a primeira baseada no método SMC-FCS, que incorpora o modelo substantivo de Cox durante o processo de imputação; e a segunda, o método MICE, que realiza a imputação sem considerar explicitamente o modelo de Cox.

5.1.2 Criação de Variáveis *dummies* para modelagem

Para a inclusão das variáveis categóricas no modelo de regressão de Cox, foi necessário criar variáveis *dummies* para que preditores não numéricos possam ser corretamente interpretados pelo modelo estatístico.

O procedimento converte uma variável categórica com k níveis em $k - 1$ novas variáveis binárias, que assumem valores de 0 ou 1. O nível que não é explicitamente transformado em uma nova variável torna-se a categoria de referência, servindo como base de comparação para os demais níveis. Todos os coeficientes e Razões de Risco (RR) para as variáveis *dummies* são, portanto, interpretados em relação a essa categoria base.

No presente estudo, o procedimento de criação de variáveis *dummies* foi aplicado às variá-

veis categóricas após a etapa de imputação, definindo-se como categorias de referência: masculino (para Sexo), 18-40 anos (para Grupo etário), cabeça, pescoço e tronco (para Localização), superficial (para Tipo Histológico), $\leq$1mm (para Índice de Breslow), ausente (para Ulceração) e não (para Presença de Mitose).

Dessa forma, a variável grupo etário, por exemplo, foi desmembrada em duas novas variáveis: `Grupo etário_2` (representando o efeito de pertencer à faixa '40-60 anos' em comparação com a faixa '18-40 anos') e `Grupo etário_3` (representando o efeito da faixa '>60 anos' em comparação com a referência). O mesmo procedimento foi aplicado a todas as demais variáveis categóricas, gerando o conjunto de dados final utilizado nos modelos de Cox.

5.1.3 Aplicação do modelo de Cox

Após a criação das variáveis *dummies*, utilizou-se o modelo de Cox com e sem a correção de Firth para obter estimativas dos parâmetros tanto nos dados de casos completos, como nos dados imputados.

Com base nos resultados da Tabela 5.2, fica evidente o quão o problema da VM pode influenciar o procedimento de estimação (no caso de dados completos, e sem aplicar a correção de Firth). Prova disso é que o erro-padrão do coeficiente associado ao fator Mitose é muito alto, evidenciando que tal fator de risco não é importante para modelar o tempo até a ocorrência da Metástase (evento de interesse). A mesma conclusão pode ser tirada analisando o elevado valor para a probabilidade de significância (Valor-p=0,997). Apesar desse comportamento, fatores que não estão diretamente associados ao problema da VM apresentam estimativas variando dentro dos padrões esperados. Os resultados mostram ainda que, indivíduos do sexo feminino são 62,7% menos susceptíveis a desenvolverem a metástase do que os do sexo masculino. De igual forma, a localização do câncer tem alguma relação com a intensidade de ocorrência da falha (83,1% de chance, quando localizado nos membros). Interpretação similar pode ser feita para os demais fatores de riscos.

As estimativas obtidas por meio da imputação múltipla demonstram maior precisão. Isso é

evidenciado pelos erros padrão, que, mesmo sem a aplicação da correção de Firth, os resultados foram consistentemente menores para todas as variáveis nos modelos com dados imputados em comparação com a análise de casos completos, indicando uma redução na variabilidade das estimativas. Diferentemente da análise de casos completos, que descarta observações com qualquer dado ausente. A abordagem por imputação utiliza o conjunto de dados no qual todas as unidades experimentais apresentam observações completas nas covariáveis, resultando em um maior número de observações para a estimação do modelo.

Ao analisar os valores-p percebe-se que as variáveis como Sexo (Feminino) e Espessura de Breslow (>4mm) são preditores significantes da sobrevida para qualquer um dos casos observados. Ademais, variáveis como Localização (Membros) e Espessura de Breslow (1,01-4mm), que não eram estatisticamente significantes na análise de casos completos (Valores-p igusia a 0,218 e 0,062), passaram a ser significativas após a imputação dos dados (Valores-p iguais a 0,019 e 0,018). A força do Índice de Breslow como preditor de um pior prognóstico é corroborada pelos resultados da Tabela 5.2, pois, suas categorias de maior espessura apresentam RR elevadas com RR=2,457 (para a faixa de 1,01-4mm) e RR=3,513 (para a faixa $>$4mm).

Além do Índice de Breslow, a localização nos membros e a presença de ulceração também se revelaram serem importantes para modelar o tempo até a ocorrência da metástase. Desta forma, os modelos com dados imputados forneceram uma visão mais confiável da magnitude dos efeitos de cada variável. A comparação dos resultados entre a análise de casos completos e as imputações por MICE e SMC-FCS forneceu um indício de que o problema da VM foi solucionado. Pois, a magnitude das estimativas do coeficiente associado a covariável Mitose, reduziu significativamente, mesmo que tal fator não tenha se tornado importante para modelar a distribuição dos tempos de falha. Por exemplo, o RR reduziu para 6,475 com um EP de 1,508. Esse resultado mostra que, indivíduos com mitose apresentam tem maior chance de desenvolver a metástase quando comparado com o nível de referência (ausência de mitose).

Tabela 5.2: Resultados do ajuste do modelo de Cox para dados completos e imputados (com, e sem a correção de Firth). Onde, Coef.: denota a estimativa para o coeficiente de regressão, EP : é o erro-padrão, e RR : denota o risco relativo.

Variável	**Casos Completos** ($n = 215$)							
	Sem correção				**Com correção**			
	Coef,	EP	Valor-p	RR	Coef,	EP	Valor-p	RR
Sexo								
Feminino	-0,992	0,434	0,022	0,371	-0,973	0,428	0,023	0,378
Grupo Etário								
40-60 anos	0,501	0,609	0,411	1,650	0,425	0,591	0,472	1,529
>60 anos	0,896	0,625	0,152	2,451	0,809	0,606	0,182	2,245
Localização								
Membros	0,605	0,491	0,218	1,831	0,599	0,487	0,219	1,820
Acral	0,242	0,579	0,676	1,274	0,282	0,575	0,624	1,325
Tipo Histológico								
Nodular	1,066	0,478	0,026	2,903	1,013	0,476	0,033	2,755
Acral	0,106	0,720	0,883	1,112	0,110	0,706	0,876	1,116
Índice de Breslow								
1,01-4mm	1,508	0,807	0,062	4,517	1,360	0,744	0,068	3,895
>4mm	1,841	0,860	0,032	6,305	1,698	0,806	0,035	5,463
Ulceração								
Presente	0,862	0,419	0,040	2,369	0,839	0,419	0,045	2,315
Presença de Mitose								
Presente	18,399	5861,597	0,997	97851205,258	1,868	1,508	0,215	6,475
	Imputação por MICE ($n = 514$)							
	Sem correção				**Com correção**			
	Coef,	EP	Valor-p	RR	Coef,	EP	Valor-p	RR
Sexo								
Feminino	-0,722	0,206	0,000	0,486	-0,717	0,206	0,000	0,488
Grupo Etário								
40-60 anos	-0,274	0,245	0,264	0,761	-0,274	0,245	0,264	0,761
>60 anos	0,263	0,254	0,299	1,301	0,259	0,253	0,307	1,296
Localização								
Membros	0,598	0,255	0,019	1,819	0,599	0,254	0,018	1,821
Acral	0,434	0,307	0,157	1,543	0,440	0,306	0,151	1,552
Tipo Histológico								
Nodular	0,998	0,334	0,003	2,713	0,987	0,333	0,003	2,683
Acral	0,647	0,423	0,126	1,910	0,647	0,421	0,124	1,910
Índice de Breslow								
1,01-4mm	0,912	0,384	0,018	2,490	0,899	0,382	0,018	2,457
>4mm	1,270	0,447	0,005	3,560	1,256	0,446	0,005	3,513
Ulceração								
Presente	0,322	0,339	0,342	1,380	0,315	0,337	0,350	1,371
Presença de Mitose								
Presente	1,399	372,455	0,997	4,052	0,746	1,094	0,495	2,109

Tabela 5.3: Resultados do ajuste do modelo de Cox para dados completos e imputados (com, e sem a correção de Firth). Onde, Coef.: denota a estimativa para o coeficiente de regressão, EP : é o erro-padrão, e RR : denota o risco relativo.

Variável	**Imputação por SMC-FCS** ($n = 514$							
	Sem correção				**Com correção**			
	Coef,	EP	Valor-p	RR	Coef,	EP	Valor-p	RR
Sexo								
Feminino	-0,852	0,203	0,000	0,427	-0,847	0,202	0,000	0,429
Grupo Etário								
40-60 anos	-0,267	0,258	0,301	0,766	-0,267	0,258	0,302	0,766
>60 anos	0,213	0,250	0,403	1,237	0,209	0,249	0,403	1,232
Localização								
Membros	0,609	0,239	0,011	1,839	0,610	0,239	0,011	1,841
Acral	0,477	0,296	0,102	1,612	0,484	0,296	0,102	1,623
Tipo Histológico								
Nodular	0,571	0,310	0,066	1,770	0,571	0,312	0,067	1,770
Acral	0,143	0,358	0,690	1,153	0,155	0,356	0,663	1,168
Índice de Breslow								
1,01-4mm	1,094	0,332	0,001	2,985	1,083	0,332	0,001	2,953
>4mm	1,641	0,377	0,000	5,158	1,633	0,375	0,000	5,121
Ulceração								
Presente	0,664	0,315	0,035	1,942	0,663	0,314	0,035	1,940
Presença de Mitose								
Presente	0,580	0,904	0,521	1,786	0,548	0,865	0,526	1,730

Por outro lado, após a aplicação dos métodos de imputação (Tabelas 5.2 e 5.3), as estimativas tornaram-se mais estáveis. Com o método MICE (com correção), o RR para Mitose foi reduzido para 2,109 (EP = 1,094), enquanto que, com o SMC-FCS, o RR foi de 1,730 (EP = 0,865). Essa redução tanto na estimativa de risco quanto na sua incerteza corrobora a eficácia da imputação para corrigir às distorções causadas pela VM. Ademais, o valor elevado para o erro-padrão (EP=372,455) associado ao fator Mitose (no caso sem correção), é consequência da influência de estimativas obtidas em dois conjuntos de dados. Nossa suspeita é que provavelmente, tais conjuntos de dados apresentaram o problema da VM.

Dentre os 50 bancos de dados imputados, pode-se verificar como ficariam as curvas de Kaplan-Meier utilizando o primeiro banco de dados de cada método, com enfoque na variável

Mitose. Tendo em conta o comportamento das curvas de sobrevivência para o fator mitose (vide na Figura 1.1), é possível observar que, após a imputação dos dados (veja a Figura 5.2), os resultados evidência que o problema da verossimilhança monótona foi contornado, o que corrobora com as observações feitas na análise da Tabela 5.2. A estabilidade e a separação entre as curvas para os grupos com e sem mitose em ambos os gráficos sugerem que os dados imputados estão livres da separação observada na análise de casos completos. Isso demonstra que a imputação não apenas permitiu o ajuste do modelo, mas também gerou cenários plausíveis onde a relação entre a mitose e a sobrevida pode ser estimada de forma finita e estável, como visto na tabela de resultados. Para visualizar as curvas de Kaplan Meier de todas as variáveis do primeiro banco de dados, verificar apêndice (Figuras B.1 e B.2).

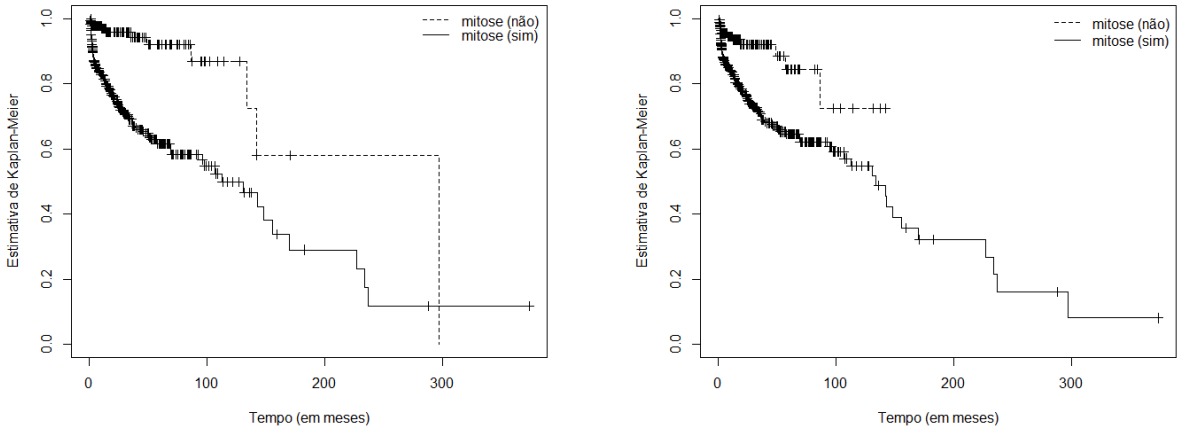

Figura 5.2: Curvas de Kaplan-Meier para a variável Mitose, geradas a partir de um único dataset representativo após imputação pelos métodos SMC-FCS (à esquerda) e MICE (à direita).

Capítulo 6

Considerações finais

O presente trabalho tratou sobre o desafio dos dados ausentes em análise de sobrevivência, um problema agravado pela ocorrência de instabilidades numéricas como a Verossimilhança Monótona (VM). deste modo, o objetivo foi investigar a aplicação e a robustez de técnicas de imputação múltipla como uma solução para garantir a validade das análises em um cenário clínico real e em estudos de simulação controlados.

O estudo de simulação revelou que a eficácia da imputação múltipla é robusta sob baixas proporções de dados faltantes, mas seu desempenho é sensível a cenários com alta ausência de dados. Na aplicação em dados reais, percebeu-se a superioridade da imputação sobre a análise de casos completos: a abordagem solucionou o problema da verossimilhança monótona, que inviabilizava a análise de casos completos, e aumentou o poder estatístico do modelo, revelando fatores de risco que antes não eram estatisticamente significantes.

As principais contribuições desta dissertação são, portanto, a demonstração prática de que a análise de casos completos pode levar a conclusões menos confiáveis em dados de sobrevivência com separação; a validação da imputação múltipla como ferramenta essencial para corrigir tais instabilidades e aumentar o poder estatístico; e a caracterização dos limites de desempenho do método MICE em cenários de simulação extremos.

Para trabalhos futuros, uma vertente de pesquisa seria a exploração de diferentes abordagens

para a construção dos modelos de imputação. Os métodos utilizados nesta dissertação baseiam-se em grande parte em modelos de regressão que assumem uma forma específica para a relação entre as variáveis.

Conclui-se que a imputação múltipla constitui uma abordagem metodológica superior e necessária para análises de sobrevivência na presença de dados ausentes, garantindo a obtenção de resultados mais robustos, precisos e clinicamente confiáveis.

Apêndice A

Resultados Adicionais

Tabela A.1: Resultados do cenário 2 (conjunto de dados contaminados com diferentes proporções de $NA's$). Sendo eMC : a estimativa de Monte Carlo; $epMC$: erro-padrão de Monte Carlo; VRM : vício relativo médio; $rEQM$: raiz quadrada do erro quadrático médio e $pCob$: probabilidade de cobertura.

Caso 1: 10% de valores faltantes										
	eMC		VRM		rEQM		epMC		pCob	
n	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
50	-6,041	0,663	-2,021	-0,010	8,254	0,267	1651,233	0,273	0,952	0,967
200	-5,446	0,664	-1,723	-0,009	7,186	0,112	657,422	0,117	0,948	0,954
500	-5,380	0,658	-1,690	-0,017	6,846	0,070	393,462	0,072	0,952	0,959
1000	-5,129	0,658	-1,565	-0,019	6,459	0,049	256,798	0,050	0,943	0,947
Caso 2: 40% de valores faltantes										
	eMC		VRM		rEQM		epMC		pCob	
n	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
50	-6,477	0,573	-2,238	-0,144	7,385	0,296	2381,380	0,321	0,948	0,971
200	-4,992	0,610	-1,496	-0,090	5,515	0,137	786,906	0,139	0,917	0,947
500	-4,722	0,621	-1,361	-0,072	5,028	0,089	447,192	0,085	0,914	0,938
1000	-4,333	0,626	-1,166	-0,066	4,588	0,070	280,194	0,060	0,907	0,903

Tabela A.2: Métricas de qualidade da imputação MICE (Caso 1: 10% e Caso 2: 40% de NA), para o cenário 2. Sendo, ACM: a acurácia média, EAM : o erro absoluto médio, $dpAC$: o desvio padrão da acurácia, e $dp.rEQM$: o desvio padrão do erro quadrático médio.

10% de valores faltantes					
	x_1		x_2		
n	ACM	dpAC	EAM	rEQM médio	dp.EQM
50	0,688	0,104	1,092	1,303	0,231
200	0,896	0,037	1,038	1,285	0,112
500	0,955	0,017	1,030	1,285	0,070
1000	0,977	0,008	1,029	1,287	0,051
40% de valores faltantes					
	x_1		x_2		
n	ACM	dpAC	EAM	rEQM médio	dp.EQM
50	0,681	0,044	1,129	1,397	0,151
200	0,888	0,011	1,050	1,312	0,070
500	0,952	0,004	1,035	1,297	0,043
1000	0,975	0,002	1,030	1,290	0,031

Apêndice B

Gráficos de curvas de Kaplan Mayer

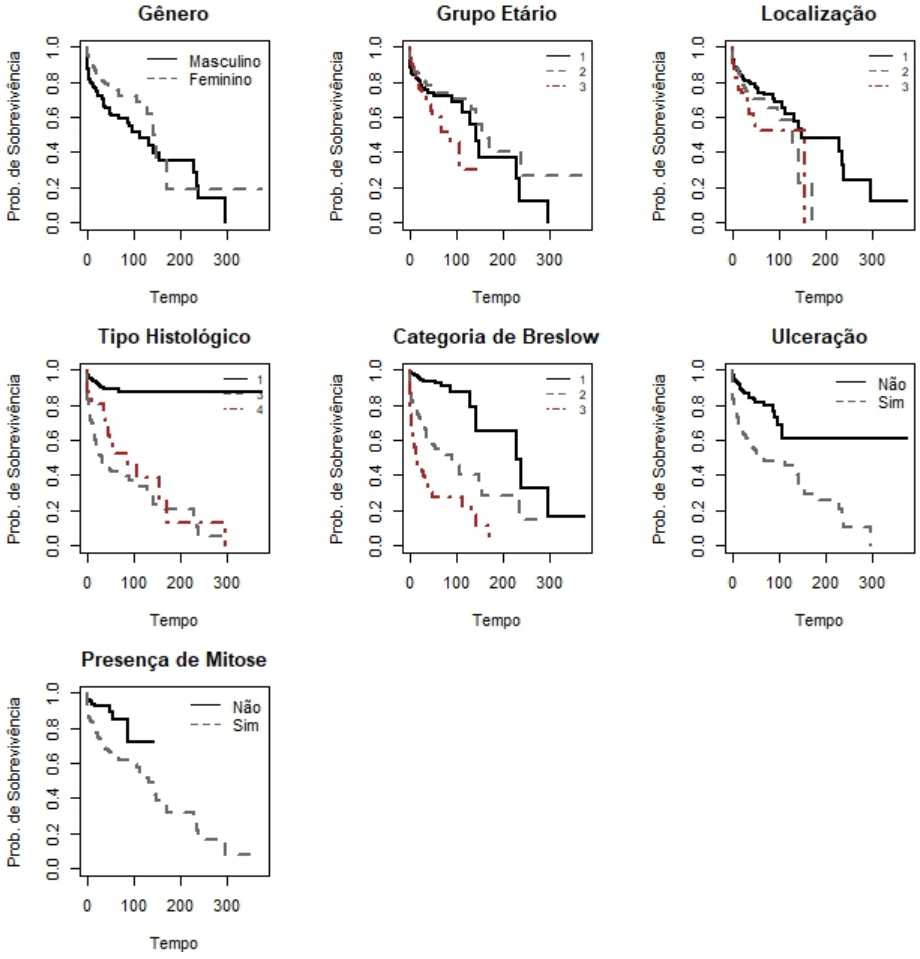

Figura B.1: Curva de Kaplan-Meier para a variável Mitose, gerada a partir de um único dataset representativo após imputação pelo método SMC-FCS.

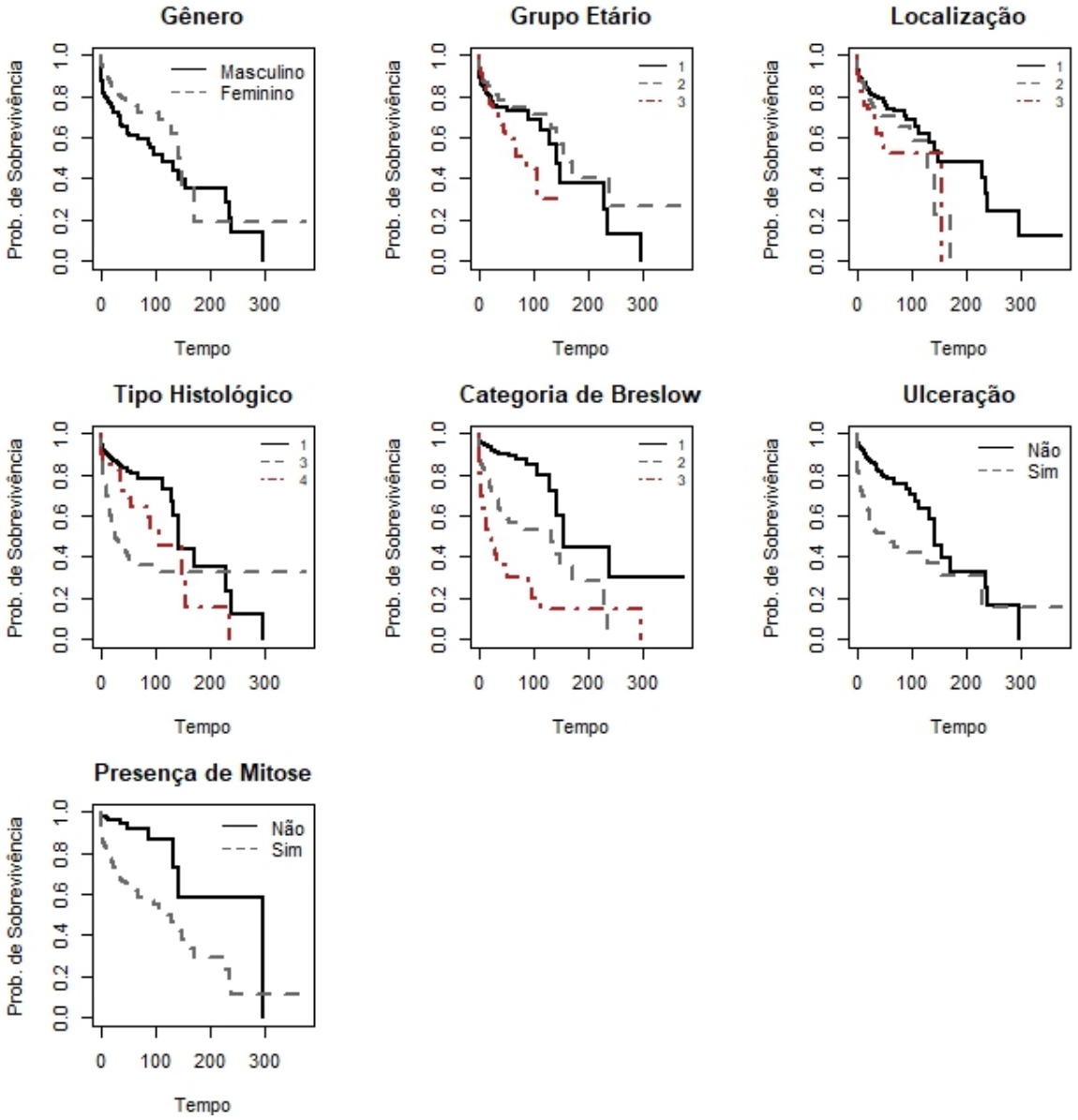

Figura B.2: Curva de Kaplan-Meier para a variável Mitose, gerada a partir de um único dataset representativo após imputação pelo método MICE.

Referências

Abd ElHafeez, Samar et al. (2021). "Methods to Analyze Time-to-Event Data: The Cox Regression Analysis". *Oxidative medicine and cellular longevity* 2021.1, p. 1302811.

Ali, A.M. et al. (2011). "Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer". *British Journal of Cancer* 104, pp. 693–699.

Allison, P. D. (2001). *Missing Data*. 1st. Thousand Oaks, CA: Sage.

Almeida, F. M. et al. (2018). "Prior specifications to handle the monotone likelihood problem in the Cox regression model". *Statistics and Its Interface* 11.4, pp. 687–698.

Almeida, Frederico M et al. (2022). "Modified score function for monotone likelihood in the semiparametric mixture cure model". *Biometrical Journal* 64.3, pp. 635–654.

Almeida, Frederico M et al. (2023a). "Bayesian solution to the monotone likelihood in the standard mixture cure model". *Statistica Neerlandica* 77.3, pp. 365–390.

Almeida, Frederico Machado et al. (2021). "Firth adjusted score function for monotone likelihood in the mixture cure fraction model". *Lifetime Data Analysis* 27, pp. 131–155.

Almeida, Gutemberg Ferreira De et al. (2023b). "IMUNOTERAPIA NO TRATAMENTO DO CÂNCER DE PELE: INIBIDORES DE CHECKPOINTS NO COMBATE AO MELANOMA". *Revista interdisciplinar em saúde*. URL: https://api.semanticscholar.org/CorpusID:258422099.

Alves, K. L. Freitas (2009). "Análise de sobrevivência de bancos privados no Brasil". Disponível em https://www.teses.usp.br/teses/disponiveis/18/18140/tde-

28102009-103529/publico/KarinaLumena_de_FreitasAlves.pdf. Tese de dout. Universidade de São Paulo.

Andersen, Per Kragh et al. (1985). "A Cox regression model for the relative mortality and its application to diabetes mellitus survival data". *Biometrics*, pp. 921–932.

Arce, Paolo M et al. (2014). "Is sex an independent prognostic factor in cutaneous head and neck melanoma?" *The Laryngoscope* 124.6, pp. 1363–1367.

Ayele, Dawit G et al. (2017). "Survival analysis of under-five mortality using Cox and frailty models in Ethiopia". *Journal of Health, Population and Nutrition* 36, pp. 1–9.

Azur, M.J. et al. (2011). "Multiple imputation by chained equations: what is it and how does it work?" *International Journal of Methods in Psychiatric Research* 20, pp. 40–49.

Bartlett, Jonathan W e Morris, Tim P (2015). "Multiple imputation of covariates by substantive-model compatible fully conditional specification". *The Stata Journal* 15.2, pp. 437–456.

Barton, Russell R e Turnbull, Bruce W (1979). "Evaluation of recidivism data: Use of failure rate regression models". *Evaluation Quarterly* 3.4, pp. 629–641.

Bergamo, G. C. (2007). "Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação". Tese (Doutorado). Piracicaba, SP, Brasil: Escola Superior de Agricultura "Luiz de Queiroz".

Brand, J. (1999). "Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets". Tese de dout. [S.l.]: Erasmus MC: University Medical Center Rotterdam, p. 26.

Breslow, N. E. e Crowley, J. (1974). "A large-sample study of the life table and product-limit estimates under random censorship". *The Annals of Statistics* 2.4, pp. 437–453.

Buhi, E. R. et al. (2008). "Out of sight, not out of mind: strategies for handling missing data". *Am J Health Behav* 32, pp. 83–92.

Buuren, S. van e Groothuis-Oudshoorn, K. (2011). "mice: Multivariate imputation by chained equations in R". *Journal of Statistical Software* 45.3, pp. 1–67.

Buuren, S. van e Oudshoorn, C. (2000). *Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual*. TNO Preventie en Gezondheid. Leiden, pp. 18, 31.

Buuren, Stef van (2018). *Flexible Imputation of Missing Data*. 2ª ed. Boca Raton: CRC Press.

Cherobin, A. C. F. P. et al. (2018). "Prognostic factors for metastasis in cutaneous melanoma". *Anais Brasileiros de Dermatologia* 93.1, pp. 19–26.

Colosimo, Enrico Antonio e Giolo, Suely Renata (2006). *Análise de Sobrevivência Aplicada*. 2ª edição. São Paulo, SP, Brasil: Editora Blucher. ISBN: 9788521204188.

Cordeiro, G. M. (1992). *Modelos de Sobrevivência: Conceitos e Aplicações*. Editora UFMG.

Costa, Júlia Carvalho et al. (2024). "ESTADIAMENTO DO MELANOMA: UMA REVISÃO NARRATIVA DE LITERATURA". *Revista Ibero-Americana de Humanidades, Ciências e Educação*. URL: https://api.semanticscholar.org/CorpusID:268561999.

Cox, D. R. (1972). "Regression models and life-tables". *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, pp. 187–220.

Cox, D. R. e Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall.

Damato, Bertil et al. (2011). "Estimating prognosis for survival after treatment of choroidal melanoma". *Progress in Retinal and Eye Research* 30.5, pp. 285–295.

Efron, B. (1977). "The efficiency of Cox's likelihood function for censored data". *Journal of the American Statistical Association* 72.359, pp. 557–565. DOI: 10.1080/01621459.1977.10480613.

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York, NY: The Guilford Press.

Ezzeddine, Wajih et al. (2015). "Cox regression model applied to pitot tube survival data". *2015 International Conference on Industrial Engineering and Systems Management (IESM)*. IEEE, pp. 168–172.

Firth, David (1993). "Bias reduction of maximum likelihood estimates". *Biometrika* 80.1, pp. 27–38. DOI: 10.1093/biomet/80.1.27.

Gelman, Andrew e Rubin, Donald B (1992). "Inference from iterative simulation using multiple sequences". *Statistical Science* 7.4, pp. 457–472.

Graham, J. W. et al. (2007). "How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory". *Prevention Science* 8, pp. 206–213.

Grover, Gurprit e Gupta, Vinay K. (2015). "Multiple imputation of censored survival data in the presence of missing covariates using restricted mean survival time". *Journal of Applied Statistics* 42.4, pp. 817–827. DOI: 10.1080/02664763.2014.986439. eprint: https://doi.org/10.1080/02664763.2014.986439. URL: https://doi.org/10.1080/02664763.2014.986439.

Gui, Jiang e Li, Hongzhe (2005). "Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data". *Bioinformatics* 21.13, pp. 3001–3008.

Haukoos, J. S. e Newgard, C. D. (2007). "Advanced Statistics: Missing Data in Clinical Research—Part 1: An Introduction and Conceptual Framework". *Acad Emerg Med* 14.7, pp. 662–668.

He, Y. (2010). "Missing data analysis using multiple imputation: getting to the heart of the matter". *Circ Cardiovasc Qual Outcomes* 3.1, pp. 98–105.

Heinze, Georg e Schemper, Michael (2006). "A comparative investigation of methods for logistic regression with separated or nearly separated data". *Statistics in Medicine* 25.24, pp. 4216–4226. DOI: 10.1002/sim.2226.

Heinze, Gunter e Schemper, Michael (2001). "A solution to the problem of monotone likelihood in Cox regression". *Biometrics* 57.1, pp. 114–119. DOI: 10.1111/j.0006-341X.2001.00114.x.

Horton, Nicholas J. e Kleinman, Ken P. (2007). "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models". *The American Statistician* 61, pp. 79–90.

Hosmer, D. W. et al. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. 2nd. Wiley-Interscience.

Hsu, Chiu-Hsieh e Yu, Mandi (2019). "Cox regression analysis with missing covariates via nonparametric multiple imputation". *Statistical Methods in Medical Research* 28.6, pp. 1676–1688. DOI: 10.1177/0962280218776988.

Hughey, Jacob J et al. (2019). "Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record". *BMC genomics* 20, pp. 1–7.

Ihwah, Azimmatul (2015). "The use of Cox regression model to analyze the factors that influence consumer purchase decision on a product". *Agriculture and Agricultural Science Procedia* 3, pp. 78–83.

Jeffreys, Harold (1946). *Theory of Probability*. Oxford University Press.

Johansen, Søren (1983). "An extension of Cox's regression model". *International Statistical Review/Revue Internationale de Statistique*, pp. 165–174.

Kalbfleisch, J. D. e Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd. John Wiley & Sons.

Kang, Hyun (2013). "The prevention and handling of the missing data". *Korean journal of anesthesiology* 64.5, pp. 402–406.

Kaplan, E. L. e Meier, P. (1958). "Nonparametric estimation from incomplete observations". *Journal of the American Statistical Association* 53.282, pp. 457–481.

Kavkler, Alenka et al. (2009). "Cox regression models for unemployment duration in Romania, Austria, Slovenia, Croatia, and Macedonia". *Romanian Journal of Economic Forecasting* 10.2, pp. 81–104.

Keiding, Niels et al. (1998). "The Cox Regression Model for Claims Data m Non-Life Insurance". *ASTIN Bulletin: The Journal of the IAA* 28.1, pp. 95–118.

Kingsley, D et al. (2008). "Cox proportional hazards regression analysis as a modeling technique for informing program improvement: Predicting recidivism in a Boys Town five-year follow-up study." *The Journal of Behavior Analysis of Offender and Victim Treatment and Prevention* 1.1, p. 82.

Klein, John P e Moeschberger, Melvin L (2006). *Survival analysis: techniques for censored and truncated data*. 2ª ed. New York: Springer.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. 2nd. Wiley.

Lima Junior, Paulo et al. (2012). "Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em física: um exemplo de uma universidade brasileira". *Revista brasileira de ensino de física* 34, p. 1403.

Linden, Andrew e Adams, Eric (2016). "Investigating the Impact of Public Health Interventions Using Multiple Imputation for Missing Data". *Epidemiology* 27.2, pp. 278–285. DOI: 10.1097/EDE.0000000000000418.

Luz, Renata Matos da et al. (2023). "Análise da variação no número dos casos de melanoma maligno e de outras neoplasias malignas da pele em todas as regiões do Brasil no período de 2015 até 2022". *Research, Society and Development*. URL: https://api.semanticscholar.org/CorpusID:259429969.

Mbona, S.V. et al. (2023). "Multiple imputation using chained equations for missing data in survival models: applied to multidrug-resistant tuberculosis and HIV data". *J Public Health Afr* 14.8, p. 2388. DOI: 10.4081/jphia.2023.2388. URL: https://doi.org/10.4081/jphia.2023.2388.

McKnight, Patrick E. et al. (2007). *Missing Data: A Gentle Introduction*. Guilford Press.

Molenberghs, G. e Verbeke, G. (2006). *Models for Discrete Longitudinal Data*. [S.l.]: Springer Science & Business Media, p. 29.

Moncada-Torres, Arturo et al. (2021). "Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival". *Scientific reports* 11.1, p. 6968.

Murphy, Susan A e Vaart, Aad W Van der (2000). "On profile likelihood". *Journal of the American Statistical Association* 95.450, pp. 449–465.

Nakano, E. Y. (2017). *Um curso de análise de sobrevivência*.

Nguyen, CD et al. (2017). "Model checking in multiple imputation: an overview and case study". *Emerging Themes in Epidemiology* 14, p. 8. DOI: 10.1186/s12982-017-0062-6.

Nunes, L. (2007). "Métodos de imputação de dados aplicados na área da saúde". Tese de Doutorado. Porto Alegre: Universidade Federal do Rio Grande do Sul.

Nunes, L. et al. (2009). "Multiple imputations for missing data: a simulation with epidemiological data". *Cad Saúde Pública* 25.2, pp. 268–278.

Nunes, L. N. et al. (2010). "Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica". *Rev Bras Epidemiol* 13.4, pp. 596–606.

Önal, Özgür et al. (2020). "Survival analysis and factors affecting survival in patients who presented to the medicaloncology unit with non-small cell lung cancer". *Turkish Journal of Medical Sciences* 50.8, pp. 1838–1850.

Papathanasiou, Dimitris et al. (2023). "Machine failure prediction using survival analysis". *Future Internet* 15.5, p. 153.

Ploner, Meinhard e Heinze, Georg (2015). "coxphf: Cox regression with Firth's penalized likelihood". *R Foundation for Statistical Computing*.

Rodrigues, Gabriela Stocco et al. (2024). "ANÁLISE EPIDEMIOLÓGICA DO MELANOMA NO BRASIL: UM ESTUDO RETROSPECTIVO (2018-2023)". *Revista CPAQV - Centro de Pesquisas Avançadas em Qualidade de Vida*. URL: https://api.semanticscholar.org/CorpusID:271509686.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley Sons.

Rubin, D. B. e Little, R. J. (2002). *Statistical Analysis with Missing Data*. Hoboken, NJ: J Wiley & Sons.

Sari, DJ et al. (2019). "Pricing life insurance premiums using Cox regression model". *AIP Conference Proceedings*. Vol. 2168. 1. AIP Publishing.

Schafer, J. L. e Graham, J. W. (2002). *Missing Data: A Gentle Introduction*. New York: Guilford Press.

Scharfstein, Daniel O et al. (2012). "On the prevention and analysis of missing data in randomized clinical trials: the state of the art". *JBJS* 94.Supplement_1, pp. 80–84.

Seaman, Shaun R et al. (2012). "What is the role of the substantive model in multivariate imputation?" *International Journal of Epidemiology* 41.2, pp. 548–560.

Sinharay, S. et al. (2001). "The use of multiple imputation for the analysis of missing data". *Psychol Methods* 6, pp. 317–329.

Thijssens, OWM e Verhagen, Wim JC (2020). "Application of extended cox regression model to time-on-wing data of aircraft repairables". *Reliability Engineering & System Safety* 204, p. 107136.

Tollenaar, Nikolaj e Van Der Heijden, Peter GM (2019). "Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes". *PloS one* 14.3, e0213245.

Tolossa, Tadesse et al. (2021). "Time to recovery from COVID-19 and its predictors among patients admitted to treatment center of Wollega University Referral Hospital (WURH), Western Ethiopia: Survival analysis of retrospective cohort study". *Plos one* 16.6, e0252389.

Turnbull, B. W. (1974). "Nonparametric estimation of a survivorship function with doubly censored data". *Journal of the Royal Statistical Society: Series B (Methodological)* 38, pp. 290–295.

Van Buuren, S. et al. (1999). "Multiple imputation of missing blood pressure covariates in survival analysis". *Statistics in Medicine* 18, pp. 681–694.

Van Buuren, Stef (2018). *Package mice: Multivariate Imputation by Chained Equations in R*.

White, I.R. e Royston, P. (2009). "Imputing missing covariate values for the Cox model". *Statistics in Medicine* 28, pp. 1982–1998.

White, I.R. et al. (2011). "Multiple imputation using chained equations: Issues and guidance for practice". *Statistics in Medicine* 30, pp. 377–399.

Wong, Ken Kwong-Kay (2011). “Using cox regression to model customer time to churn in the wireless telecommunications industry”. *Journal of Targeting, Measurement and Analysis for Marketing* 19, pp. 37–43.