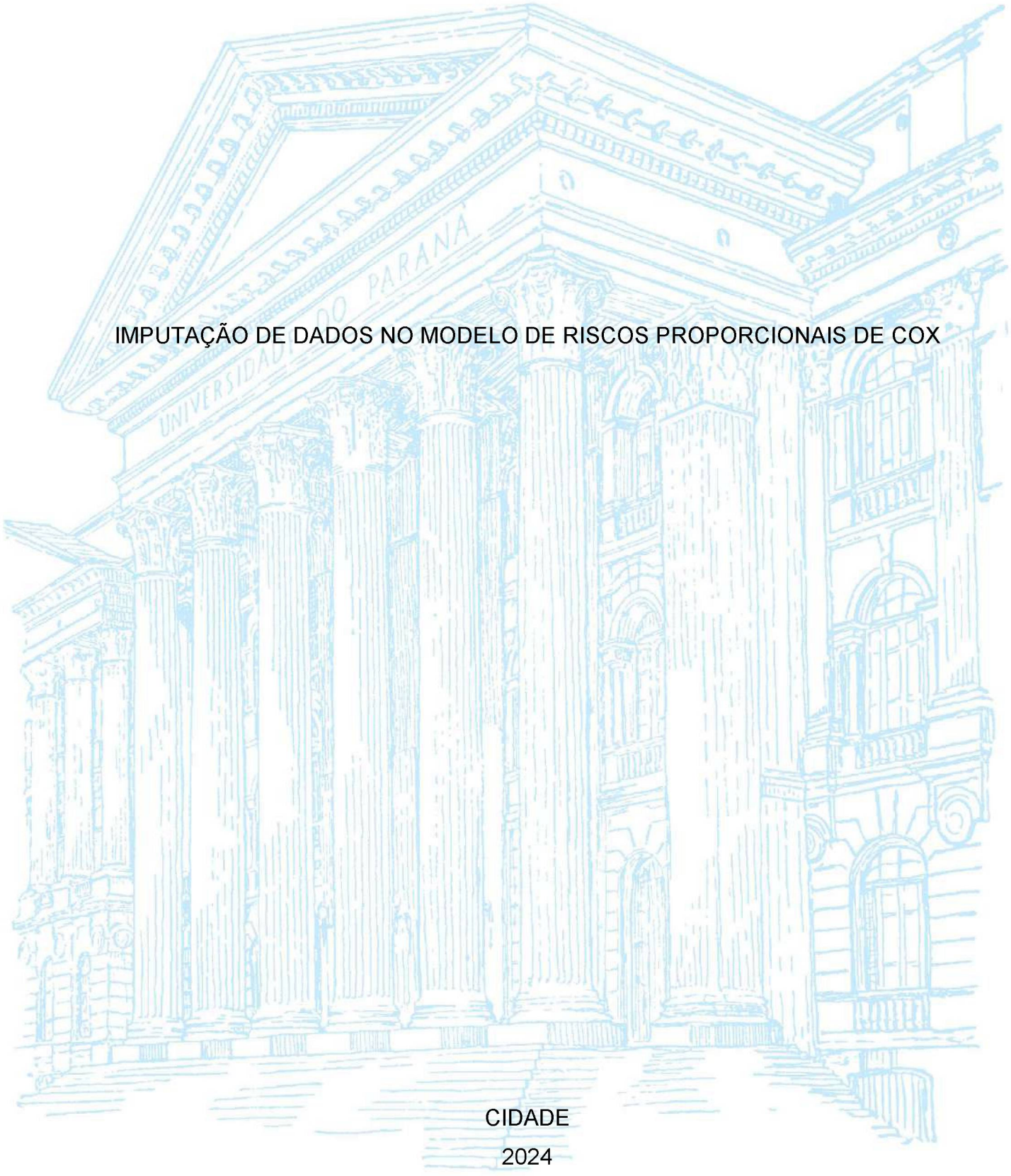
UNIVERSIDADE FEDERAL DO PARANÁ

BIANCA SALLES TRENTIN



IMPUTAÇÃO DE DADOS NO MODELO DE RISCOS PROPORCIONAIS DE COX

CIDADE

2024

BIANCA SALLES TRENTIN

IMPUTAÇÃO DE DADOS NO MODELO DE RISCOS PROPORCIONAIS DE COX

Dissertação apresentada ao curso de Pós-Graduação em Métodos Numéricos em Engenharia, Setor de Tecnologia e Ciências Exatas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Métodos Numéricos em Engenharia.

Orientador: Prof. Dr. José Luiz Padilha da Silva

CURITIBA

2024

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

Trentin, Bianca Salles
Imputação de dados no modelo de riscos proporcionais de COX / Bianca Salles Trentin. – Curitiba, 2024.
1 recurso on-line : PDF.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Métodos Numéricos em Engenharia.

Orientador: José Luiz Padilha da Silva

1. Análise de sobrevivência (Biometria). 2. Imputação múltipla (Estatística). 3. Observações faltantes (estatísticas). I. Universidade Federal do Paraná. II. Programa de Pós-Graduação em Métodos Numéricos em Engenharia. III. Silva, José Luiz Padilha da. IV . Título.

Bibliotecário: Leticia Priscila Azevedo de Sousa CRB-9/2029

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **BIANCA SALLES TRENTIN** intitulada: **IMPUTAÇÃO DE DADOS NO MODELO DE RISCOS PROPORCIONAIS DE COX**, sob orientação do Prof. Dr. JOSE LUIZ PADILHA DA SILVA, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestra está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 30 de Setembro de 2024.

Assinatura Eletrônica
01/10/2024 20:31:31.0
JOSE LUIZ PADILHA DA SILVA
Presidente da Banca Examinadora

Assinatura Eletrônica
07/10/2024 11:39:49.0
SILVANA SCHNEIDER
Avaliador Externo (UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL)

Assinatura Eletrônica
07/10/2024 15:09:55.0
CESAR AUGUSTO TACONELI
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Dedico este trabalho ao meu avô, Darci Campos Salles.

AGRADECIMENTOS

Iniciar esta jornada acadêmica foi um desafio imenso, e nada teria sido possível sem o apoio de tantas pessoas que, direta ou indiretamente, contribuíram para a realização deste trabalho. Gostaria de expressar minha profunda gratidão ao meu orientador, Professor José Luiz Padilha da Silva, por sua orientação incansável, paciência e suporte ao longo de todo o processo. Suas orientações precisas e críticas construtivas foram essenciais para o desenvolvimento desta dissertação. Aos meus pais, por estarem ao meu lado em todos os momentos, apoiando minhas escolhas e oferecendo amor e encorajamento incondicional. Vocês me ensinaram o valor do esforço e da resiliência. Ao meu avô, Darci Campos Salles, que sempre foi um exemplo de honestidade e caráter. Sua partida neste último ano deixou um vazio enorme, mas sua memória e ensinamentos me deram força para seguir em frente. Esse trabalho é, de certa forma, um tributo à pessoa incrível que ele foi. Um agradecimento especial ao meu namorado, por sua paciência, carinho e compreensão durante os altos e baixos dessa jornada. Seu apoio constante e incentivo nos momentos difíceis me deram a força necessária para seguir em frente. Aos meus amigos, que sempre estiveram prontos para oferecer apoio, seja com palavras de incentivo ou com momentos de descontração que foram essenciais para manter o equilíbrio durante os momentos de maior pressão. Agradeço à Universidade Federal do Paraná e ao Programa de Pós-Graduação em Métodos Numéricos em Engenharia pela oportunidade de realizar este trabalho. Por fim, sou grata a todas as pessoas que de alguma forma contribuíram para a realização deste trabalho. Cada gesto, por menor que tenha sido, fez toda a diferença.

"Mudanças não acontecem de uma vez; elas se realizam pouco a pouco."

Simone de Beauvoir

RESUMO

A imputação múltipla (RUBIN, 1987) é essencial em análises estatísticas quando ocorrem dados ausentes. No modelo de taxas de falhas proporcionais (COX, 1972), contudo, métodos convencionais de imputação são inadequados. Conduzimos um estudo de simulação comparando quatro abordagens de imputação múltipla para duas variáveis explicativas parcialmente observadas, considerando que a perda de dados está relacionada ao tempo de sobrevivência. Avaliamos os modelos convencionais de imputação, regressão linear (NORM) para dados contínuos e para dados binários o modelo de regressão linear logística (LOG). Para acomodar as características do modelo de taxas de falhas proporcionais optamos por uma abordagem baseada na função taxa de falha acumulada em uma aproximação do modelo de imputação proposto por White e Royston (2009) (NA), um método compatível com o modelo de análise proposto por Bartlett et al. (2015) (CONG) e o método CART, conhecido por lidar bem com não linearidades e interações. Para variáveis contínuas, o método NORM e o NA apresentaram maior viés do que o modelo de caso completo (CC). Os métodos NORM, CART e NA resultaram em baixas taxas de cobertura. No cenário das variáveis binárias, as taxas de cobertura para os métodos LOG, CART e NA aumentaram em comparação com as variáveis contínuas. Neste cenário, o método CART teve menor erro padrão, resultando em taxas de cobertura abaixo do nível nominal. Em ambos os cenários, o modelo de imputação CONG ofereceu estimativas razoáveis em comparação com os demais métodos, apresentando menor viés e taxas de cobertura alinhadas aos valores nominais. Na aplicação prática dos métodos ao conjunto de dados de pacientes com Doença de Chagas do HC-UFMG, reforçou-se a importância de uma escolha cuidadosa da técnica de imputação múltipla para garantir a robustez das inferências em estudos de sobrevivência com dados ausentes. Dessa forma, a pesquisa contribui para o avanço na utilização de técnicas de imputação múltipla em modelos de sobrevivência, apontando o método CONG como uma alternativa viável e eficiente em contextos práticos.

Palavras-chave: Análise de sobrevivência; Ausência de dados; Imputação múltipla; Modelo de Cox

ABSTRACT

Multiple imputation (RUBIN, 1987) is a fundamental technique in statistical analyses addressing missing data. However, conventional imputation methods may be insufficient in the context of the Cox proportional hazards model (COX, 1972). This study presents a simulation that compares four multiple imputation approaches for two partially observed covariates, focusing on the scenario where the missing data mechanism is associated with survival time. We evaluated traditional imputation models, specifically linear regression (NORM) for continuous variables and logistic regression (LOG) for binary variables. To align with the characteristics of the Cox model, we employed an approach based on the cumulative hazard function, as proposed by White e Royston (2009) (NA), in addition to a method compatible with the analysis framework outlined by Bartlett et al. (2015) (CONG) and the CART method, which is known for its capacity to handle non-linearities and interactions effectively. For continuous variables, the NORM and NA methods exhibited greater bias compared to the complete case (CC) model, while the NORM, CART, and NA methods demonstrated low coverage rates. In the binary variables scenario, the coverage rates for the LOG, CART, and NA methods improved relative to the continuous variables, with the CART method showing the lowest standard error, leading to coverage rates that fell below the nominal level. Across both scenarios, the CONG imputation model provided reasonable estimates, with lower bias and coverage rates that aligned closely with nominal values. In the practical application of these methods to a dataset of Chagas disease patients from HC-UFMG, the findings emphasize the importance of selecting appropriate multiple imputation techniques to ensure the robustness of inferences in survival studies with missing data. Thus, this research contributes to advancing the use of multiple imputation techniques in survival models, highlighting the CONG method as a viable and efficient alternative in practical applications.

Keywords: Survival analysis; Missing data; Multiple imputation; Proportional hazards model

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os métodos estatísticos clássicos foram desenvolvidos para acomodar matrizes de dados provenientes de estudos planejados, ou seja, nas linhas da matriz temos as observações e nas colunas as variáveis medidas para cada unidade observada. No entanto, em determinadas ocasiões alguns dados de entrada da matriz de dados não são completamente observados.

Em estudos clínicos, de acordo com Austin et al. (2020), o motivo da ausência de dados se deve a fatores variados, como por exemplo a perda de seguimento do paciente, erro de máquinas para exames, erro médico, e médicos que não solicitam certas investigações para alguns pacientes. Além destes fatores, a condição de saúde do indivíduo também pode contribuir para a ausência de dados.

De acordo com Little e Shencker (1995) o problema com dados ausentes é que eles podem levar a análises tendenciosas, estimativas menos eficientes e complicações em métodos estatísticos, ou seja, a ausência de dados pode criar vieses se as unidades com valores ausentes forem sistematicamente diferentes das unidades com dados completos. Nur et al. (2010) destacam as implicações de dados ausentes em estudos clínicos. Além do viés nas estimativas dos parâmetros no modelo de análise, podemos estar diante da perda de poder estatístico, resultando em conclusões incorretas caso os dados ausentes não sejam tratados de forma adequada.

Antes de aplicar um tratamento para os dados ausentes é necessário compreender os mecanismos de ausência. Carpenter e Kenward (2012) destacam a dificuldade em determinar o mecanismo de perda de dados, mas ressaltam a importância de analisá-lo para estabelecer a robustez das inferências. De acordo com Sainani (2015), o impacto de dados ausentes depende do motivo pelo qual os dados estão ausentes, e diferentes tipos de mecanismo de perda exigem diferentes estratégias de tratamento.

Na literatura nos deparamos com três padrões de dados ausentes: *missing completely at random* (MCAR), *missing at random* (MAR) e *missing not at radom* (MNAR). A premissa de ausência aleatória (MCAR), frequentemente utilizada, simplifica análises, mas Yi et al. (2009) alertam que nem sempre a análise é válida em todos os casos. Conforme Raghunnathan (2004), os tratamentos para dados ausentes incluem três abordagens distintas. O IPW (*Inverse Probability Weighting*), funciona atribuindo pesos a casos completos com base no inverso na probabilidade de ser um caso completo. Os pesos são usados para dar mais influência a casos completos e ajustar a exclusão de indivíduos com dados ausentes, fornecendo assim uma inferência válida para a população.

Outro método aplicado no tratamento de dados ausentes é o da máxima verossimilhança (*Maximum Likelihood* (ML)). A ML no contexto de de dados ausentes busca estimar os parâmetros do modelo a partir dos dados observados sem necessariamente preencher

ão os campos ausentes da matriz de dados. O processo de estimação no método ML, usualmente, faz a implementação do algoritmo *Expectation-Maximization* (EM). Rubin (1987), desenvolveu o paradigma da Imputação Múltipla (IM) para incorporar a incerteza da imputação de dados. Basicamente, a IM gera K conjuntos de dados imputados e as estimativas resultantes são combinadas em uma estimativa final.

Nas referências de tratamento de dados ausentes em estudos de análise de sobrevivência, podemos observar um problema comum em pesquisa clínica que é a ausência de dados. Considerando o modelo de taxas de falhas proporcionais estudos como de Carroll, Morris e Keogh (2020), analisam como os pesquisadores lidam com dados faltantes em covariáveis em estudos observacionais de tempo até evento na oncologia. Outros autores, como por exemplo Guo, Yang e Yi-Hau (2021), aplicaram métodos IM via KNN e *Random Forest* comparando os padrões de ausência MCAR, MAR e MNAR. No artigo de Hsu e Yu (2019), os autores aplicaram o método da probabilidade inversa aumentada ponderada (AIPW), o método da imputação de correspondência média preditiva (PMM) e a uma abordagem proposta de imputação múltipla não paramétrica.

Além das pesquisas citadas anteriormente, os autores Yi et al. (2009) simularam três cenários distintos de ausência de dados considerando o padrão de ausência de dados em variáveis explicativas dependentes do tempo. Na pesquisa de Yi et al. (2009), a ausência ocorria em uma variável que poderia ser de natureza contínua ou binária e o tratamento aplicado foi o IPW.

A pequisa de Silva (2023) replicou os cenários simulados por Yi et al. (2009) e aplicou métodos via IM para tratar os dados ausentes. Silva (2023), considerou o método baseado na função taxa de falha acumulada (NA) (WHITE; ROYSTON, 2009), um método compatível com o modelo de análise (CONG) (BARTLETT et al., 2015) e o método *Classification And Regression Tree*(CART) (BREIMAN, 2001). Além da comparação de métodos em cenários simulados, Silva (2023) realizou uma aplicação prática em um banco de dados de pacientes com doença de Chagas. Os dados utilizados para a análise prática são provenientes de um estudo clínico realizado pelo Hospital de Clínicas da Universidade Federal de Minas Gerais (HC-UFMG). Os pacientes que participaram desse estudo foram acompanhados de 1999 até 2019. O autor, Silva (2023), relata que em situações práticas frequentemente temos dados ausentes em múltiplas variáveis explicativas e dependentes do tempo, o que representa um desafio adicional à análise.

1.1 JUSTIFICATIVA

A presente pesquisa é similar à realizada por Silva (2023), mas com algumas alterações. Para o modelo mais simples de análise, simulamos a perda em múltiplas variáveis explicativas de natureza contínua e binária. Além disso, consideramos um padrão de ausência independente da variável resposta e das explicativas. E ainda, avaliamos uma ausência de

aproximadamente 80% em uma das covariáveis, isto porque na análise prática com a base de dados de Chagas encontramos uma variável explicativa com esta característica que não foi explorada anteriormente.

1.2 OBJETIVOS

1.2.1 OBJETIVO GERAL

Comparar os mecanismos de dados ausentes e o viés das estimativas de coeficientes no modelo de taxas de falhas proporcionais, utilizando o método de correção imputação múltipla, considerando variáveis contínuas e binárias, bem como a ausência em múltiplas variáveis explicativas.

A escolha de focar em variáveis contínuas e binárias se deve à natureza específica dos dados de doença de Chagas analisados e aos objetivos do estudo. O objetivo do estudo de simulação, foi concentrar-se em variáveis mais frequentes em estudos clínicos, que geralmente envolvem dados contínuos e binários.

1.2.2 OBJETIVOS ESPECÍFICOS

- Entender os diferentes mecanismos de dados ausentes que podem ocorrer em conjuntos de dados aplicados ao modelo de taxas de falhas proporcionais, tanto para variáveis contínuas quanto binárias.

- Avaliar o impacto dos diferentes mecanismos de dados ausentes nas estimativas de coeficientes do modelo de taxas de falhas proporcionais para variáveis contínuas e binárias.

- Aplicar o método IM para tratar o problema de dados ausentes e verificar sua eficácia na redução do viés das estimativas de coeficientes.

2 REVISÃO DE LITERATURA

2.1 ANÁLISE DE SOBREVIVÊNCIA

A análise de sobrevivência é um método estatístico amplamente utilizado em muitos campos. Nesta abordagem, a variável resposta é o tempo do evento de interesse.

> Normalmente o valor da variável aleatória é o tempo até a falha de um componente físico (mecânico ou elétrico) ou o tempo até a morte de uma unidade biológica (paciente, animal, célula, etc.)(JR.; GONG; MUÑOZ, 1998).

Em estudos clínicos, o tempo até o evento de interesse pode ser denominado como o tempo até a falha, um exemplo de falha seria a morte de determinado indivíduo sob observação. De acordo com Kleinbaum e Klein (2005), em termos de representação matemática podemos descrever o tempo de falha como T que é uma variável aleatória. Conforme Colosimo e Giolo (2006), a variável T é não-negativa e é usualmente contínua. Usualmente o tempo de falha é especificado pela função de sobrevivência ou pela função taxa de falha.

Um aspecto importante dos dados de sobrevivência é a censura, que representa observações limitadas da resposta em estudo por diversas causas. Podemos definir o tempo de censura C como uma variável aleatória independente de T, portanto, o tempo observado t pode ser definido como o mínimo entre o tempo de falha e o tempo de censura. As fórmulas a seguir podem ser encontradas em Colosimo e Giolo (2006),

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

A função indicadora de sobrevivência é representada por δ e pode se expressa por

$$\delta = \begin{cases} 1 & \text{se } T \leq C, \\ 0 & \text{se } T > C. \end{cases} \tag{2.1}$$

A função de sobrevivência pode ser definida como a probabilidade de uma observação não falhar até um determinado tempo observado t e pode ser expressa da seguinte maneira,

$$S(t) = P(T > t),\ t \geq 0.$$

E a função de distribuição acumulada é definida por,

$$F(t) = 1 - S(t).$$

A função taxa de falha pode ser definida como,

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

Como consequência, temos a função taxa de falha acumulada,

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

2.2 ESTIMADOR DE KAPLAN-MEIER

O estimador de Kaplan-Meier (KM) é usado para estimar a função de sobrevivência em estudos de tempo até o evento de interesse. Basicamente, o estimador calcula a probabilidade de um indivíduo sobreviver além de um determinado tempo t, considerando a censura.

Seja $t_1 < t_2 < \ldots < t_k$ os k tempos distintos e ordenados de falha, d_j é o total de falhas em $(t_{j-1}, t_j]$ e n_j o total de indivíduos sob o risco em t_{j-1}, $j = 1, \ldots, k$, podemos escrever o estimador KM da seguinte maneira,

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right). \tag{2.2}$$

2.3 ESTIMADOR PARA TAXA DE FALHA ACUMULADA

A função de taxa de falha acumulada, $\Lambda(t)$, desempenha um papel crucial na análise de dados de sobrevivência. Apesar de não possuir uma interpretação direta ela pode ser importante na avaliação taxa de falha instantânea que pode ser representada por $\lambda(t)$.

De acordo com Colosimo e Giolo (2006), na estimação não paramétrica, em específico, a função taxa de falha acumulada possui um estimador com propriedades ótimas, o que a torna uma ferramenta importante na modelagem de dados de sobrevivência. Todavia, a estimação direta de $\Lambda(t)$ pode ser trabalhosa, exigindo métodos mais avançados e abordagens estatísticas mais refinadas. Uma maneira de realizar a estimativa da função taxa de falha acumulada é através do estimador de Nelson-Aalen, que é baseado na função de sobrevivência expressa por

$$S(t) = \exp\{-\Lambda(t)\}. \tag{2.3}$$

O estimador de Nelson-Aalen, apresentado por Nelson (1972) e retomado por Aalen (1978), é muito utilizado na análise de sobrevivência para estimar a função de risco cumulativa. Desta forma o estimador é dado por

$$\tilde{\Lambda}(t) = \sum_{j:t_j<t} \left(\frac{d_j}{n_j}\right).$$

A variância pode ser estimada da seguinte forma

$$\widehat{\text{Var}}(\tilde{\Lambda}(t)) = \sum_{j:t_j<t} \left(\frac{d_j}{n_j^2} \right).$$

As componentes d_j e n_j são definidas igual ao estimador KM (Equação 2.2)

2.4 MODELO DE RISCOS PROPORCIONAIS DE COX

Comumente, as pesquisas na área da medicina incluem covariáveis que podem influenciar a duração da sobrevivência. A literatura relacionada à análise de sobrevivência apresenta duas categorias de modelos: os paramétricos e os semiparamétrico.

De acordo com Kleinbaum e Klein (2005), um modelo paramétrico assume uma forma funcional específica da função de risco ou sobrevivência. Já a modelagem semiparamétrica permite a estimativa não-paramétrica da função taxa de falha, o que resulta em maior flexibilidade.

Iremos nos concentrar na classe de modelos semiparamétricos. O modelo de taxas de falha proporcionais proposto por Cox (1972). Conforme Box-Steffensmeier e Jones (2004), os resultados da pesquisa original de Cox foram amplamente aplicados a problemas em todos os domínios científicos. Além disso, a lógica do modelo de taxas de falhas proporcionais é simples e elegante. Considere o seguinte modelo para um tempo de falha contínuo T e um vetor de covariáveis X. Queremos fazer inferências sobre um vetor de parâmetros $\boldsymbol{\theta}$, logo o modelo pode ser expresso por

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp\{\theta' \mathbf{X}\}.$$

Onde $\lambda(t|\mathbf{X})$ é função taxa de falha no tempo t, dado $\mathbf{X}$, e $\lambda_0(t)$ é uma função de taxa de falha baseline especificada e é comum a todas as observações.

2.4.1 VEROSSIMILHANÇA PARCIAL

Levando em consideração a parte não-paramétrica, $\lambda_0(t)$, do modelo de taxas de falhas proporcionais na função de verossimilhança, o método usual para estimação dos parâmetros é inadequado. Uma alternativa viável é restringir a formulação da função de verossimilhança com base no entendimento do histórico de falhas e censuras, o que leva à remoção da função de perturbação da verossimilhança. Esse procedimento é chamado de método da verossimilhança parcial.

Segundo Klein e Moeschberger (2003), assumimos que a censura é não-informativa porque o tempo de censura não está relacionado ao risco ou à probabilidade de ocorrência da falha, ou seja, estamos assumindo que o fato de um indivíduo ser censurado não carrega informações sobre a probabilidade de que ele teria experimentado o evento de interesse.

Considerando uma amostra de n indivíduos que possui k números de falhas e que $k \leq n$ sob a suposição de que não há vínculo entre os tempos dos eventos, denotamos os tempos de eventos diferentes e ordenados como $t_1 < t_2 < \cdots < t_k$. De acordo com Colosimo e Giolo (2006), consideramos a probabilidade condicional de um indivíduo falhar em t_j, conhecendo os que estão sob risco em t_j. Podemos definir o conjunto dos indivíduos $\mathcal{R}(t_j)$ sob o risco no tempo t_j. Além disso, consideramos que $\mathbf{x}_j$ é o vetor de covariáveis do indivíduo que falhou em t_j. Desta forma, Podemos expressar a verossimilhança parcial como,

$$\mathrm{L}(\boldsymbol{\theta}) = \prod_{j=1}^{k} \frac{\exp\{\boldsymbol{\theta}'\mathbf{x}_j\}}{\sum_{\ell \in \mathcal{R}(t_j)} \exp\{\boldsymbol{\theta}'\mathbf{x}_\ell\}}. \tag{2.4}$$

Nota-se que o numerador depende apenas da informação do indivíduo que tem o evento de interesse, enquanto o denominador utiliza informações sobre todos os indivíduos que ainda não sofreram o evento, incluindo indivíduos que podem ser censurados. Podemos dizer que o indivíduo estar "sob risco" significa que ele ainda está sendo observado no estudo e não sofreu o evento de interesse até um determinado tempo t_k, ou seja, podemos dizer que o indivíduo não foi censurado ou não experimentou o evento até o tempo t_k.

As estimativas de máxima verossimilhança parciais são encontradas resolvendo o sistema de equações escore provenientes da verossimilhança parcial, ou seja, o gradiente da função de verossimilhança parcial em relação aos parâmetros $\boldsymbol{\theta}$. Basicamente, resolvemos um sistema de equações que iguala o vetor escore a zero, o que nos permite encontrar os elementos do vetor $\boldsymbol{\theta}$ que maximizam a verossimilhança. Como estamos estimando θ_p parâmetros, logo temos p equações não-lineares para resolver, que correspondem aos índices de p variáveis explicativas no modelo de taxas de falhas proporcionais. As estimativas podem ser obtidas através da técnica de Newton-Raphson.

2.5 AUSÊNCIA DE DADOS

De acordo com Falcaro et al. (2015), o contexto de análise de sobrevivência requer a necessidade de utilizar técnicas de imputação de dados robustas, que são capazes de lidar com não linearidades, interações complexas entre variáveis e fornecer estimativas menos viesadas. Além disso, essas técnicas devem garantir que a cobertura empírica das estimativas esteja próxima dos níveis nominais.

Conforme Pigott (2010), a ausência de dados pode causar vários problemas na análise, incluindo: resultados tendenciosos, precisão reduzida nas estimativas de parâmetros, incapacidade de generalização, relações estatísticas não confiáveis e aplicabilidade limitada do modelo de análise. Diante do modelo de taxas de falhas proporcionais. Além desses problemas, podemos ter o inflacionamento do erro tipo I que é quando rejeitamos a hipótese nula quando ela é verdadeira e este é descrito por Guo, Yang e Yi-Hau (2021).

Quando nos deparamos com a ausência de dados faz sentido considerar análises que tratam este problema. De acordo com Sainani (2015), não existe uma solução única para lidar com dados ausentes e a estratégia ideal depende do desenho do estudo, dos objetivos da análise e do padrão de dados faltantes.

2.5.1 MECANISMOS E PADRÕES DE AUSÊNCIA DE DADOS

Segundo Little e Rubin (2002), o mecanismo de dados ausentes diz respeito à relação entre a ausência e os valores das variáveis na matriz de dados. É caracterizado pela distribuição condicional da matriz indicadora de ausência.

Carpenter e Kenward (2012) destacam que é difícil saber, de fato, qual o mecanismo de perda de dados. No entanto, alguns mecanismos são viáveis, pois apresentam consistências em relação aos dados observados. Na prática, deseja-se analisar as informações através de mecanismos únicos, a fim de estabelecer a robustez das inferência frente à incerteza sobre o mecanismo de ausência.

Neste sentido, o interesse está nas implicações práticas da ausência de dados tanto para estimar parâmetros quando para fazer inferência para a população de interesse. Carpenter e Kenward (2012) ressalvam que, a ausência de dados, os métodos computacionais apenas conduzem a inferências válidas sob pressupostos específicos. Uma estrutura formal para descrever esse comportamento não só é necessária como fundamental.

Considere uma amostra de n indivíduos de uma população infinita. Seja $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,p})^T$, representando p variáveis a serem coletadas na i-ésima unidade, com $i = 1, \ldots, n$. Denotamos por $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ um conjunto de p parâmetros populacionais.

Para cada unidade $i = 1, \ldots, n$ denotamos $\mathbf{Y}_{i,O}$ o subconjunto observado na base de dados e $\mathbf{Y}_{i,M}$ o subconjunto que está ausente. Assim, para diferentes indivíduos, $\mathbf{Y}_{i,O}$ e $\mathbf{Y}_{i,M}$ são subconjuntos diferentes das variáveis. Se nenhum dado estiver faltando, $\mathbf{Y}_{i,M}$ estará vazio. Seja a função indicadora de ausência $R_{i,j}$, para cada unidade $i = 1, \ldots, n$ e as variáveis $j = 1, \ldots, p$. Definimos $R_{i,j} = 1$ se $Y_{i,j}$ é observado e $R_{i,j} = 0$ se $Y_{i,j}$ é ausente. Podemos reescrever a função indicadora de ausência de maneira transposta $\mathbf{R}_i = (R_{i1}, \ldots, R_{ip})^T$. Como $\mathbf{R}_i$ define o mecanismo de ausência, matematicamente temos que

$$P(\mathbf{R}_i|\mathbf{Y}_i). \tag{2.5}$$

A expressão (2.5) representa a probabilidade condicional da função indicadora de ausência $\mathbf{R}_i$ dado o conjunto de dados $\mathbf{Y}_i$.

2.5.1.1 *MISSING COMPLETELY AT RANDOM* (MCAR)

A ausência completamente aleatória (MCAR) ocorre quando a falta de um valor específico em uma variável não tem relação com os dados observados ou não observados.

Algebricamente podemos representar este mecanismo da seguinte forma,

$$P(\mathbf{R}_i|\mathbf{Y}_i) = P(\mathbf{R}_i).$$

Dado que, em situações de MCAR, a probabilidade de dados faltantes não está vinculada aos valores, os dados observados tornam-se representativos da população. Um exemplo deste tipo de ausência é quando uma pesquisa coleta dados sobre renda, mas algumas respostas não são registradas por causa de um erro técnico aleatório do sistema de coleta. Neste caso, a ausência dos dados de renda não está relacionada com o valor da renda e nem com outras variáveis observadas no estudo.

2.5.1.2 *MISSING AT RANDOM* (MAR)

Definimos os dados como *Missing At Random* (MAR) quando, condicionados aos dados observados, a distribuição de probabilidade de R_i é independente dos dados não observados,

$$P(\mathbf{R}_i|\mathbf{Y}_i) = P(\mathbf{R}_i|\mathbf{Y}_{i,O}).$$

Um exemplo de ausência do tipo MAR pode ser em pesquisas nas quais pessoas com certos perfis (como idade) tem mais probabilidade de não responder a perguntas sobre saúde, mas a ausência das respostas não depende diretamente dos valores de saúde ausentes. Além do mecanismo de ausência, temos o conceito de ausência não-ignorável. Segundo Molenberghs e Kenward (2007), a ausência não-ignorável refere-se a uma situação em que a falta de dados não é aleatória e não pode ser ignorada na análise. Neste sentido, de acordo com Carpenter e Kenward (2012) ,os dados ausentes são sistematicamente diferentes dos dados observados, e a probabilidade de ausência depende dos valores das variáveis no conjunto de dados. Um exemplo disso seria a ausência do tipo MAR quando a perda depende da resposta observada.

2.5.1.3 *MISSING NOT AT RANDOM* (MNAR)

Em um mecanismo MNAR, a probabilidade de uma observação estar ausente está relacionada ao valor subjacente, e essa dependência persiste mesmo após considerar os dados observados. Algebricamente,

$$P(\mathbf{R}_i|\mathbf{Y}_i) \neq P(\mathbf{R}_i|\mathbf{Y}_{i,O}).$$

Embora, em certos contextos, seja mais plausível considerar o MNAR em vez do MAR, a análise sob o MNAR é substancialmente mais desafiadora. Conforme demonstrado por Carpenter e Kenward (2012), isso se deve ao fato de que, no MAR, as distribuições condicionais de variáveis parcialmente observadas, dados os valores completamente observados, permanecem iguais entre unidades que possuem e não possuem os dados observados.

Para este caso, as inferências geralmente requerem a especificação do modelo correto para o mecanismo de dados faltantes e suposições distributivas para as covariáveis faltantes, ou ambos.

Supondo que a variável resposta seja a renda de determinado indivíduo, um exemplo de MNAR ocorre quando indivíduos com alta renda optam por não responder a uma pesquisa de renda porque não querem revelar seus ganhos. Nesse caso, a ausência de dados está diretamente relacionada aos valores não observados da renda, ou seja, pessoas com rendas mais altas são mais propensas a não responder. Dessa forma, se considerarmos que os indivíduos com renda mais alta são diferentes dos indivíduos com renda mais baixa, estamos diante de uma ausência não ignorável.

2.6 TRATAMENTO PARA DADOS AUSENTES

A análise de caso completo (CC) é a técnica que exclui qualquer linha com observações faltantes no conjunto de dados. Dentre os tratamentos utilizados para dados ausentes esta é a estratégia mais simples e mais comumente usada. Todavia, pode levar à perda de poder, precisão estatística e viés nas estimativas.

A imputação única é bastante usual e substitui os valores faltantes por uma estimativa plausível. Entretanto, reduz a variabilidade e pode levar a erros padrão e valores p artificialmente baixos. De acordo com Sainani (2015) a imputação única é adequada para um pequeno número de valores ausentes.

Para variáveis com maior percentual de ausência podemos considerar o método da Imputação Múltipla (IM). O método de tratamento via IM trata a incerteza através de um processo em várias etapas. Primeiramente, cria-se K conjuntos de dados "completos", cada um com valores substituídos de forma diferente para refletir a incerteza sobre os dados ausentes. Em seguida, cada conjunto é analisado separadamente. Os resultados dessas análises são combinados usando formulações específicas que consideram tanto a variabilidade dentro de cada conjunto de dados quanto a variabilidade entre os diferentes conjuntos imputados. Este processo permite que a incerteza associada à imputação seja incorporada nas estimativas finais.

Há ainda outra forma de tratar os dados, de acordo com Seaman e White (2011) o método IPW ajusta a análise de dados incompletos atribuindo pesos aos casos completos com base no inverso da probabilidade de serem completos. Os pesos estimados são usados na análise para fornecer inferências válidas para a população, assumindo que o modelo de imputação esteja corretamente especificado. Outro método para tratar dados ausentes é o da ML. De acordo com Pigott (2010), esse método estima os parâmetros de um modelo com dados faltantes maximizando a verossimilhança dos dados observados. Quando não há solução fechada para essa maximização, o algoritmo EM pode ser utilizado. O algoritmo EM consiste em duas etapas iterativas. A primeira etapa, chamada *E-step*,

estima o valor esperado dos dados ausentes, dados os parâmetros atuais do modelo. Em outras palavras, utiliza-se a distribuição condicional dos dados observados com base nos parâmetros estimados na iteração anterior. O objetivo é calcular o que seriam os dados faltantes, supondo que os parâmetros do modelo são verdadeiros.

A segunda etapa, *M-step*, utiliza as estimativas obtidas na *E-step* para maximizar a função de verossimilhança, recalibrando os parâmetros do modelo. Isso é feito ajustando a média e a covariância das estimativas. O processo itera até que as estimativas se estabilizem. Esse método foca nos parâmetros dos dados observados e requer técnicas especializadas para calcular os erros padrão, devido à complexidade associada à incerteza dos dados não observados. Além disso, o ML difere da verossimilhança usual, que considera apenas dados completamente observados. Na presença de dados faltantes, o ML não ignora a informação potencialmente útil que os dados ausentes podem fornecer, utilizando o algoritmo EM para lidar com essa incerteza.

2.6.1 IMPUTAÇÃO MÚLTIPLA

A imputação múltipla tem sido amplamente adotada como uma estratégia para lidar com a ausência de dados durante a análise estatística. Segundo White e Royston (2009), diferentemente dos métodos de imputação mais simples, a IM pode oferecer a vantagem de produzir inferências que refletem de forma precisa a incerteza decorrente dos dados faltantes.

Diante da suposição de que o modelo de imputação está corretamente especificado, de acordo com Zhou, Eckert e Tierney (2001), a imputação múltipla na análise de sobrevivência melhora estimativas de parâmetros ao lidar com dados faltantes e permite imputar vários valores faltantes.

Goretzko, Heumann e Bühner (2019) destacam que a escolha do método de imputação pode impactar a precisão dos valores imputados e análises estatísticas subsequentes. Portanto, os autores recomendam a avaliação cuidadosa de diferentes métodos de imputação que podem ser necessários para garantir inferências válidas de dados com valores ausentes.

Na IM, os valores ausentes são imputados várias vezes por meio de simulações, criando múltiplos conjuntos de dados completos com imputações diferentes para os valores faltantes. Cada valor ausente é substituído por um conjunto de $K > 1$ valores plausíveis retirados de sua distribuição preditiva (SCHAFER; OLSEN, 1998).

A distribuição preditiva refere-se à distribuição dos possíveis valores que podem ser imputados, com base nas informações disponíveis nas variáveis observadas. Esta abordagem é fundamentada na estatística bayesiana, e pode ser representada pela expressão $f(\boldsymbol{Y}_M|\boldsymbol{Y}_O,\boldsymbol{R})$. A incerteza sobre os valores ausentes é modelada utilizando a informação que está presente nos dados. Em termos práticos, um modelo estatístico é ajustado aos dados disponíveis, que leva em consideração a relação entre as variáveis observadas. A

distribuição preditiva é derivada dessa modelagem. Ao gerar múltiplas imputações, a técnica permite que diferentes valores sejam amostrados da distribuição preditiva.

A discrepância entre as K imputações espelha a incerteza associada à previsão dos valores ausentes com base nos valores observados. Após a condução de análises em cada um dos K conjuntos de dados, os resultados, incluindo estimativas e erros-padrão, são agregados de acordo com as diretrizes delineadas por Rubin (1987), resultando em estimativas globais e erros padrão que incorporam a incerteza relacionada aos dados faltantes.

Fundamentando-se em Carpenter e Kenward (2012), supondo que estamos diante de um problema convencional para um modelo estatístico com um vetor de parâmetros $\boldsymbol{\theta}$, com dimensão $(p \times 1)$. Podemos denominar o modelo de análise como modelo *substantivo* e é assumido que se os dados não são ausentes, ou seja, os dados são completos, podemos fazer inferências válidas sobre algum ou todos os elementos de $\boldsymbol{\theta}$. Um estimador consistente para $\boldsymbol{\theta}$ pode ser obtido pela equação,

$$\sum_{i=1}^{n} \mathbf{U}_i(\hat{\boldsymbol{\theta}}; \mathbf{Y}) = U(\hat{\boldsymbol{\theta}}; \mathbf{Y}) = \mathbf{0}. \tag{2.6}$$

A expressão (2.6) representa a função escore, cada vetor $U(\hat{\boldsymbol{\theta}}; \mathbf{Y})$ representa a derivada da função de verossimilhança em relação aos parâmetros do modelo, avaliada em um conjunto de dados imputados. Podemos calcular um estimador consistente para a matriz de variância e covariância de θ, denotado $\mathbf{Var}(\boldsymbol{\theta})$. Um estimador consistente para θ para os dados observados (incompleto) pode ser obtido pela equação,

$$E_{f(\boldsymbol{Y}_M|\boldsymbol{Y}_O,\boldsymbol{R})}\{\boldsymbol{U}(\hat{\boldsymbol{\theta}}; \boldsymbol{Y}_O, \boldsymbol{Y}_M)\} = 0. \tag{2.7}$$

A equação (2.7), a notação $E_{f(\boldsymbol{Y}_M|\boldsymbol{Y}_O,\boldsymbol{R})}$ indica que estamos calculando a esperança da função de distribuição condicional para os dados ausentes, dado os dados observados e a matriz indicadora de ausência. A equação (2.7) estabelece que a esperança da função escore, levando em conta a incerteza dos dados ausentes, deve ser igual a zero. Isso implica que o estimador de $\hat{\boldsymbol{\theta}}$ é um ponto crítico na função de verossimilhança, ou seja, onde a função atinge um ponto de máximo. Sob a suposição de MAR não é necessário condicionar o indicador de ausência, então $f(\mathbf{Y}_M|\mathbf{Y}_O, R) = f(\mathbf{Y}_M|\mathbf{Y}_0)$, seguindo a definição do mecanismo MAR isto se deve ao fato de que a probabilidade de que os dados estejam ausentes depende apenas dos dados observados e não dos dados ausentes em si. Basicamente $\mathbf{R}$ não fornece mais informações sobre os valores ausentes, em outras palavras a suposição MAR permite ignorar $\mathbf{R}$ ao condicionar a distribuição dos dados ausentes em função dos observados.

Para uma explicação mais aprofundada, o processo de imputação múltipla implica na substituição dos valores ausentes pelos valores correspondentes de K conjuntos de imputação, resultando em K conjuntos de dados completos. Se denotamos $\hat{\boldsymbol{\theta}}_k$ como a

estimativa de $\boldsymbol{\theta}_k$ e $\hat{\boldsymbol{V}}_{\boldsymbol{k}}$ como a matriz de covariância associada ao k-ésimo conjunto de dados completo, a estimativa da IM para $\boldsymbol{\theta}$ é obtida calculando a média simples das estimativas e as expressões podem ser encontradas em Carpenter e Kenward (2012).

$$\hat{\boldsymbol{\theta}}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\theta}}_k. \tag{2.8}$$

A matriz de covariância média dentro da imputação:

$$\hat{\boldsymbol{W}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{V}}_k. \tag{2.9}$$

A matriz de covariância entre imputações de $\hat{\boldsymbol{\theta}}_k$:

$$\hat{\boldsymbol{B}} = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{MI})(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{MI})^T. \tag{2.10}$$

e estimativa do covariância $\hat{\boldsymbol{\theta}}_{MI}$ é dada por

$$\hat{\boldsymbol{V}}_{MI} = \hat{\boldsymbol{W}} + \left(1 + \frac{1}{K}\right)\hat{\boldsymbol{B}}. \tag{2.11}$$

De forma resumida, as inferências nos dados imputados seguem as regras de Rubin expressa pelas equações (2.8), (2.9), (2.10) e (2.11) para combinar as estimativas de parâmetros após a imputação múltipla. Para cada conjunto de dados completos (após a imputação), estimamos os parâmetros de interesse, denotados por $\hat{\boldsymbol{\theta}}_k$ para o cada conjunto k, e matriz de variância associada $\hat{\mathbf{V}}_k$. A inferência é realizada combinando essas estimativas $\hat{\boldsymbol{\theta}}_k$ e suas variações. A estimativa final de θ é obtida pela média dos parâmetros (Equação (2.8)).

A matriz de variância $\hat{\mathbf{V}}_{MI}$ é calculada combinando a variância dentro das imputações $\hat{\mathbf{W}}$ com a variância entre as imputações $\hat{\mathbf{B}}$, a fim de capturar a incerteza gerada pelo processo de imputação. Para isso, utilizam-se as equações (2.9) e (2.10). A estimativa final da matriz de variância é dada pela equação (2.11). Esse processo incorpora a incerteza associada aos dados ausentes para a estimativa final de $\boldsymbol{\theta}$.

A distribuição de referência para as estimativas $\hat{\boldsymbol{\theta}}_{MI}$ segue uma distribuição t com graus de liberdade ajustados pela quantidade de imputações K e pela variabilidade entre as imputações. A estatística de teste T pode ser referenciada a uma distribuição t com ν graus de liberdade,

$$T = \frac{\hat{\theta}_{MI} - \theta_0}{\sqrt{\hat{V}_{MI}}}.$$

O termo θ_0 representa a hipótese nula para os parâmetros de interesse θ. Em testes de hipóteses, θ_0 é o valor que se assume verdadeiro sob a hipótese nula H_0. No contexto de

IM θ_0 é o valor esperado do parâmetro θ sob a hipótese nula. Os graus de liberdade ν são ajustados pela fórmula

$$\nu = (K-1)\left(\frac{1+\frac{1}{K}\hat{W}}{B+W}\right)^2,$$

refletindo a incerteza adicional introduzida pelo processo de imputação múltipla. Assim, os intervalos de confiança e os testes de hipótese podem ser conduzidos utilizando a distribuição t com os graus de liberdade ajustados.

3 MÉTODOS DE IMPUTAÇÃO

3.1 *IMPUTAÇÃO MÚLTIPLA: ACOMODANDO O MODELO SUBSTANTIVO* (CONG)

De acordo com Carpenter e Kenward (2012), o relacionamento não linear pode complicar as imputações. Se considerarmos que apenas uma pequena proporção de valores faltantes está presente nas covariáveis, a análise por casos completos (CC) geralmente é a opção mais apropriada, pois o impacto da exclusão de dados ausentes é minimizado. No entanto, em situações onde esse cenário não se aplica, é necessário utilizar métodos de imputação para garantir uma análise consistente e robusta.

O modelo proposto por Bartlett et al. (2012) (CONG), modifica o algoritmo FCS (Fully Conditional Specification) padrão para garantir que os modelos de imputação univariados sejam compatíveis com o modelo de análise. Basicamente, o FCS utiliza modelos condicionais, onde para cada variável com dados ausentes é definido um modelo que imputa os valores faltantes com base nas demais variáveis observadas. Cada variável ausente é tratada individualmente e seu valor é atualizado iterativamente, levando em consideração as outras variáveis como variáveis explicativas no modelo condicional.

O processo de imputação ocorre em ciclos sucessivos, onde as variáveis são imputadas uma de cada vez, e a cada iteração os valores imputados são atualizados. Isso é repetido até que as imputações se estabilizem e as distribuições dos parâmetros convirjam.

A incorporação do método de aceitação e rejeição, proposta por Bartlett et al. (2015), na amostragem é pela necessidade de gerar amostras de uma distribuição alvo que pode ser difícil de amostrar diretamente. Todavia, no contexto da amostragem por aceitação e rejeição utilizamos uma distribuição proposta, da qual é fácil gerar amostras, e então é aplicado um critério de aceitação baseado na relação entre a distribuição alvo e a distribuição proposta. Esse critério permite comparar uma amostra gerada da distribuição proposta com a densidade da distribuição alvo, rejeitando ou aceitando com base em uma probabilidade calculada.

Cabe ressaltar que esta estratégia é necessária quando a densidade da distribuição alvo não pertence a uma família paramétrica simples ou quando é difícil representar diretamente amostras dela, como ocorre em muitos modelos substantivos complexos. A rejeição irá ocorrer quando a amostra não se adequar ao critério estabelecido pela distribuição alvo, garantindo que as amostras sigam a distribuição correta.

O primeiro passo em cada ciclo do algoritmo é ajustar o modelo substantivo aos valores observados e imputados atuais. Neste caso, estamos diante do modelo de riscos proporcio-

nais. O resultado do ajuste nos retornará estimativas de máxima verossimilhança parciais $\hat{\theta}$ e a matriz de variância e covariância associada $\hat{\boldsymbol{\Omega}}$. Então realizamos a aproximação: (i) gerar $\boldsymbol{\theta}$ da normal multivariada $N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Omega}})$ e então com os valores de $\boldsymbol{\theta}$ estimamos a função taxa de falha basal $h_0(t)$ e a função taxa de falha acumulada $H_0(t)$.

Para a variável X_1, condicional as outras, defina a distribuição proposta $g(.)$ e considere a proposta $X_{i1}^{\star}$ para o i-ésimo indivíduo ausente em X_1, no valor atual de $\boldsymbol{\theta}$. A função $g(.)$ representa a proposta para os dados ausentes e deve ser especificada de acordo com a classe da variável que foi perdida. As expressões a seguir são descritas por Carpenter e Kenward (2012), considere um indivíduo censurado, $\delta_i = 0$, a probabilidade de aceitação da amostra é dada por:

$$S(T_i|X_{i,1}^{\star}, X_{i,2}; \boldsymbol{\theta}).$$

Para um indivíduo não censurado, $\delta_i = 1$, a probabilidade de aceitação é dada por,

$$H_0(t)exp\left[1 + (X_{i,1}^{\star}\,\theta_1 + X_{i,2}\,\theta_2) - H_0(t)exp\{X_{i,1}^{\star}\theta_1 + X_{i,2}\theta\}\right].$$

No modelo proposto, $X_{i,1}^{\star}$ é o valor de $g(.)$ que foi proposto e aceito pelo método de aceitação e rejeição. Conforme a sugestão de Carpenter e Kenward (2012) para garantir a convergência precisamos obter um número suficiente de iterações, normalmente em torno de 10 e é necessário que o algoritmo atinja uma distribuição estacionária.

As estimativas de $H(t)$ são condicionalmente geradas com base nas covariáveis imputadas, uma vez que o *software* utilizado, R, gera estimativas condicionais e não marginais de $H(t)$, o que é consistente com a forma de imputação feita. Isso assegura que os parâmetros estimados levem em consideração as dependências condicionais do modelo substantivo. Entretanto, em determinadas ocasiões $H(t)$ pode não depender das covariáveis. Diante do risco acumulado marginal $H(t)$ irá refletir o risco acumulado para a população em geral, sem condicionar às covariáveis. Assim, diferentemente das estimativas condicionais fornecidas pelo R, onde $H(t)$ é ajustado de acordo com os valores das preditoras, utilizando uma forma marginal de $H(t)$, o que pode ser interpretado como uma estimativa global do risco acumulado ao longo do tempo, independentemente das características individuais representadas pelas covariáveis. Nesse contexto, as imputações realizadas devem ser compatíveis com essa natureza marginal de $H(t)$, já que o foco não está nas variações do risco acumulado com base nas covariáveis, mas sim no comportamento global do risco ao longo do tempo

Segundo Bartlett et al. (2012), por construção as desvantagens do método incluem a possibilidade de incompatibilidade entre o modelo de imputação e o modelo substantivo, ou seja, o modelo de imputação precisa ser compatível com o modelo de análise. Caso o modelo seja incompatível, o resultado será estimativas assintoticamente enviesadas dos parâmetros do modelo de análise, especialmente se este inclui efeitos não lineares ou interações.

Além disso, a especificação de um modelo conjunto para as covariáveis parcialmente observadas é desafiadora, especialmente quando há uma combinação de variáveis contínuas e discretas, o que pode comprometer a precisão dos resultados.

3.2 IMPUTAÇÃO BASEADA NA TAXA DE FALHA ACUMULADA (NA)

O método de imputação baseado na taxa de falha acumulada é uma aproximação do modelo correto de imputação, proposto por White e Royston (2009).

Este método estima a função taxa de falha acumulada a partir dos dados observados. Basicamente, é realizada uma regressão das covariáveis ausentes sobre a estimativa de risco, incluindo a interação com o tempo de sobrevivência, de forma que a imputação considere a relação entre as covariáveis, o tempo de sobrevivência e o evento de interesse.

Considere X_1 e X_2 como variáveis ausentes e Z como uma variável completa. Sob o modelo de taxas de falhas proporcionais, a distribuição condicional de X_1 e X_2, dado Z e o tempo de sobrevivência, é aproximada por

$$\begin{aligned}\log\{f(X_1, X_2|T,\delta, Z)\} = \log\{f(X_1, X_2|Z)\} + \delta[\log\{h_0(T)\} + (\theta_1 X_1 + \theta_2 X_2 + \theta_3 Z)] \\ - H_0(T)\exp(\theta_1 X_1 + \theta_2 X_2 + \theta_3 Z).\end{aligned} \tag{3.1}$$

Para variáveis binárias X_1 e X_2, podemos aplicar uma aproximação por série de Taylor de primeira ordem, válida quando a variância de Z é pequena, resultando em:

$$\log\{P(Y_1 = 1 \mid T, \delta, Y_2, Y_3)\} \approx \zeta_0 + \zeta_1 X_2 + \zeta_2 Z + \zeta_3 \delta + \zeta_4 H_0(T) \times X_2 + \zeta_6 H_0(T) \times Z. \tag{3.2}$$

Para $\zeta_0, \ldots, \zeta_6$ constantes. Isto é, a equação (3.2) é aproximadamente uma regressão logística que considera Z como covariável completa, δ como indicadora de censura, e $H_0(T)$ como estimativa de risco acumulada.

Quando X_1 e X_2 são contínuas, em particular $X_1, X_2|Z \sim N(\alpha_0 + \alpha_1 Z, \sigma^2)$, da equação (3.1) temos a seguinte aproximação:

$$\log\{f(X_1, X_2|T, \delta, Z)\} \approx \frac{((X_1 + X_2) - \alpha_0 - \alpha_1 Z)^2}{2\sigma^2} + \delta(\theta_1 X_1 + \theta_2 X_2) - H_0(T)\, e^{(\theta_1 X_1 + \theta_2 X_2 + \theta_3 Z)}. \tag{3.3}$$

As funções apresentadas estão relacionadas ao modelo de imputação, pois descrevem a distribuição condicional de X_1, X_2 dado Z, o tempo de sobrevivência T e a indicadora de censura δ. Essas funções são a base para a construção do modelo de imputação no contexto de riscos proporcionais utilizando a função taxa de falha acumulada basal $H_0(t)$. A ideia principal do método é imputar valores faltantes de X_1, X_2 de forma que as relações entre as covariáveis e o tempo até o evento de interesse sejam consideradas.

Quando X_1, X_2 são binários, ajusta-se a regressão logística conforme a equação (3.2), em que X_1, X_2 são modelados em função de Z, δ, e $H_0(T)$. Nesse caso, a aproximação por séries de Taylor é usada para simplificar a relação entre X_1, X_2 e essas covariáveis. Se X_1, X_2 é contínuo, o modelo segue a equação (3.3), que assume que X_1, $X_2|Z$ segue uma distribuição normal, levando em consideração os termos de interação com δ e $H_0(T)$. Todavia, não podemos fazer suposições de distribuições a respeito dos dados que serão imputados, neste caso a (3.3) que assume normalidade está considerando apenas a natureza contínua das variáveis que são ausentes.

Após ajustar o modelo de imputação, os valores são utilizados para preencher os campos ausentes de X_1, X_2. No caso binário, a probabilidade predita de $X_1, X_2 = 1$ é usada para gerar as imputações, enquanto no caso contínuo, os valores de X_1, X_2 são imputados a partir da distribuição normal ajustada.

Caso $H_0(T)$ seja desconhecido, ele deve ser previamente estimado. O estimador de Nelson-Aalen (Equação 2.3) pode ser utilizado para esse fim, de modo a calcular a função de risco acumulado com base nos dados observados antes de realizar a imputação.

Segundo os autores, White e Royston (2009), este método de imputação é útil em modelos de riscos proporcionais, mas apresenta limitações em certos contextos. Ele assume que os efeitos das covariáveis são pequenos e que a incidência cumulativa é baixa, o que pode introduzir viés quando essas condições não são atendidas. A incidência cumulativa mede a probabilidade de ocorrência de um evento ao longo do tempo para um grupo específico de indivíduos, ela é calculada como o número de novos casos de um evento em um período de tempo dividido pelo número de indivíduos em risco no início desse período. Outros problemas neste método, descritos por White e Royston (2009), em situações com múltiplas covariáveis interativas ou não lineares, o método se torna mais complexo e pode não capturar adequadamente essas interações. A precisão da estimativa da função de risco acumulado $H_0(T)$ é importante, e imprecisões nessa estimativa podem comprometer a imputação.

3.3 IMPUTAÇÃO MÚLTIPLA VIA ÁRVORES DE CLASSIFICAÇÃO E REGRESSÃO (CART)

As árvores de classificação e regressão (CART) proposta por Breiman (1984) são uma classe popular de algoritmos de aprendizado de máquina. Os modelos CART buscam preditores e pontos de corte nos preditores usados para dividir a amostra. Os pontos de corte dividem a amostra em sub amostras mais homogêneas. O processo de divisão é repetido em ambas as subamostras, de modo que uma série de divisões define uma árvore binária. A variável alvo pode ser discreta (árvore de classificação) ou contínua (árvore de regressão). Podemos resumir o algoritmo de imputação sob um modelo de árvore usando

o *bootstrap* com os seguintes passos:

1. Sorteie uma amostra de *bootstrap* $(\dot{y}_{obs}, \dot{X}_{obs})$ de tamanho n_1 a partir de (y_{obs}, X_{obs}).
2. Ajuste $\dot{y}_{obs}$ por $\dot{X}_{obs}$ usando um modelo de árvore $f(X)$.
3. Faça previsões para os n_0 nós terminais g_j a partir de $f(X_{mis})$.
4. Construa n_0 conjuntos Z_j com todos os casos no nó g_j, cada um contendo d_j doadores candidatos.
5. Sorteie aleatoriamente um doador i_j de Z_j para $j = 1, \ldots, n_0$.
6. Calcule as imputações $\dot{y}_j = y_{i_j}$ para $j = 1, \ldots, n_0$.

Podemos interpretar a imputação CART considerando que y_{obs} é variável ausente e X_{obs} é o vetor de covariáveis que contém o tempo de falha T, a indicadora de censura C e outras covariáveis, como por exemplo Z. O primeiro passo consiste em sortear uma amostra *bootstrap* dos dados. Essa amostra é caracterizada por outra notação no caso $(\dot{y}_{obs}, \dot{X}_{obs})$. No segundo passo, onde o modelo de árvore é adaptado aos dados de treinamento $(\dot{y}_{obs}, \dot{X}_{obs})$, existe uma margem considerável para escolhas. A função $f(X_{mis})$ refere-se à função preditiva que é ajustada usando os dados observados e que, posteriormente, é aplicada a um conjunto de dados com valores ausentes.

É vantajoso configurar a árvore de forma que haja um número fixo de casos em cada nó, por exemplo, 5 ou 10. Isso porque a composição dos grupos de doadores será diferente em diferentes repetições do procedimento de *bootstrap*, o que ajuda a incorporar a incerteza associada à amostragem na construção da árvore.

De acordo com Burgette e Reiter (2010) o CART é menos eficiente do que os modelos paramétricos e pode causar descontinuidades nos limites de partição. No entanto, isso não é uma grande preocupação ao usar o CART para imputações. Em estudos de sobrevivência, apesar de incluir a variável C como preditora, esse método não foca em lidar com a censura presente nos dados. Silva (2023) relata que o método subestima a incerteza da imputação, resultando em baixas taxas de cobertura. Além disso, a *pool* de doadores pode não refletir a incerteza de amostragem correta, especialmente em tamanhos de amostra pequenos ou quando o modelo de imputação contém variáveis fortemente relacionadas à ausência de dados. Isso leva a intervalos de confiança que apresentam probabilidades de cobertura inferiores aos valores nominais. Todavia, o método pode ser aplicado tanto em modelos não paramétricos quanto em modelos paramétricos, pode ser usado com variáveis contínuas e categóricas, dependentes ou independentes. Além disso, o CART lida bem com não linearidades e variáveis correlacionadas.

4 ESTUDO DE SIMULAÇÃO

Um estudo de simulação foi conduzido a fim de comparar os métodos de imputação sob diferentes padrões de ausência no modelo de taxas de falhas proporcionais. Especificamente, exploramos os mecanismos MCAR, MAR e MNAR. Levamos em consideração a natureza binária e contínua das variáveis com ausência de dados. Além disso, comparamos o desempenho dos métodos de regressão linear (NORM e LOG), CONG, NA e CART. O objetivo principal da simulação é investigar a robustez das inferências diante de diferentes métodos de imputação, padrão de ausência e natureza da covariável.

Considere o seguinte modelo de análise para um tempo de falha contínuo T e um vetor de covariáveis $\mathbf{V} = (X_1, X_2, Z)$. Queremos fazer inferências sobre um vetor de parâmetros $\boldsymbol{\theta}$.

$$\lambda(t|\mathbf{V}) = \lambda_0(t)\ exp(\theta_1 X_1 + \theta_2 X_2 + \theta_3 Z) \tag{4.1}$$

O vetor de parâmetros é $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T = (1,1,1)^T$ e $\lambda_0(t)$ foi considerada constante e igual a 1 para todas observações. Consideramos um cenário em que $X_1 \sim N(0,1)$ é parcialmente observada e as demais covariáveis, $X_2 \sim N(0,1)$ e $Z \sim Bin(n, 0.5)$, são completamente observadas. Além disso, o cenário considera que as variáveis X_1 e X_2 são parcialmente observadas e $X_1, X_2 \sim N(0,1)$. A variável Z é completamente observada e $Z \sim Bin(n, 0.5)$.

A escolha de $\lambda_0(t)$ constante e igual a 1 é com base na adequação para o cenário inicial da simulação. Apesar de $\lambda_0(t)$ restritiva por não refletir a complexidade de cenários mais práticos, essa alternativa se torna viável por não precisamos nos preocupar com a modelagem de $\lambda_0(t)$.

Além do cenário para covariáveis contínuas, simulamos um cenário em que $X_1 \sim Bin(n,0.5)$ é parcialmente observada e as demais covariáveis, $X_2 \sim Bin(n,0.5)$ e $Z \sim Bin(n, 0.5)$, são completamente observadas. Analogamente ao cenário contínuo, consideramos que as variáveis X_1 e X_2 são parcialmente observadas e X_1 e $X_2 \sim Bin(n,0.5)$. A variável Z é completamente observada e $Z \sim Bin(n, 0.5)$. R é uma função indicadora que assume 0 quando a variável é não observada e 1 quando é observada (Tabela 1).

Tabela 1 – Configurações dos Cenários Simulados

Configurações	**Padrão de Ausência**	**Variável ausente**	**Probabilidade de Observação**	**Viés**
Cenário 1	MAR	X_1	$P(R_{X_1} = 1) = (1 + \exp\{-T\})^{-1}$	Sim
Cenário 2	MAR	X_2	$P(R_{X_2} = 1) = (1 + \exp\{-T + 2(X_1 - 0.5) + Z\})^{-1}$	Sim
Cenário 3	MCAR	X_1	$P(R_{X_1} = 1) = 0.60$	Não
Cenário 4	MAR MNAR	X_1, X_2	$P(R_{X_1} = 1) = (1 + \exp\{-T\})^{-1}$, $P(R_{X_2} = 1) = (1 + \exp\{-T + 2(X_1 - 0.5) + Z\})^{-1}$	Sim
Cenário 5	MCAR MNAR	X_1, X_2	$P(R_{X_1} = 1) = 0.60$, $P(R_{X_2} = 1) = (1 + \exp\{-T + 2(X_1 - 0.5) + Z\})^{-1}$	Sim

Para cada cenário apresentado, consideramos os modelos de imputação,

1. Modelo usual de imputação via regressão linear, para variáveis de natureza contínua (NORM) e binárias (LOG).
2. Aproximação do modelo de imputação utilizando taxa de falha acumulada (NA).
3. Modelo compatível com o de análise que leva em conta o método de aceitação e rejeição para imputar um valor proposto (CONG).
4. Modelo de árvores, conhecido por lidar bem com não linearidades e interações (CART).

Além disso, para efeitos de comparação, iremos nos referir ao modelo ajustado como **FULL**, supondo que tivéssemos o banco de dados completamente observado. O Caso Completo (**CC**) é o modelo ajustado com a remoção das linhas onde há valores ausentes. Para avaliar os resultados, utilizamos as seguintes métricas de avaliação em N que indica o tamanho da sua amostra de dados.

1. **Viés**: Seja N o número total de amostras geradas via Monte Carlo (MC). Cada estimativa $\hat{\theta}_i$ é obtida em cada uma das simulações, e o viés mede a diferença média entre as estimativas e o verdadeiro valor do parâmetro

$$\text{Viés}(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_{\text{real}}).$$

2. **Erro padrão (EP)**: Esta métrica avalia a incerteza associada à estimativa dos coeficientes $\hat{\theta}_i$, obtidos via máxima verossimilhança parcial. Para calcular o EP, é necessário primeiro obter a matriz de informação de Fisher, que reflete a curvatura da função de verossimilhança no ponto das estimativas dos coeficientes.

$$\mathbf{I}(\hat{\theta}) = \frac{\partial^2 L(\hat{\boldsymbol{\theta}})}{\partial\theta\partial\theta'}.$$

   A função $L(\hat{\boldsymbol{\theta}})$ é definida na Equação 2.4. A matriz de variância e covariância do estimador $\hat{\theta}$ pode ser expressa como a inversa da matriz de informação:

   O EP de cada coeficiente $\hat{\theta}$ é obtido extraindo a raiz quadrada do valor correspondente na diagonal principal da matriz de variância e covariância

$$EP(\hat{\theta}_i) = \sqrt{\text{Var}(\hat{\theta}_i)}.$$

3. **Desvio Padrão (DP)**: O desvio padrão empírico mede a variabilidade das estimativas dos coeficientes obtidas a partir das simulações de Monte Carlo (MC). Ele é calculado com base na distribuição das estimativas para cada tamanho amostral N.

$$DP(\hat{\theta}_i) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(\hat{\theta}_i - \overline{\hat{\theta}_i})^2}.$$

4. **Cobertura Empírica (CE)**: Verifica se o verdadeiro valor do parâmetro θ_{real} está dentro do intervalo de confiança,

$$\text{LI} \leq \theta_{\text{real}} \leq LS.$$

Os limites inferior (LI) e superior (LS) são obtidos utilizando $LI = \hat{\theta} - t_{\nu,0.975} \times SE$ e $LS = \hat{\theta} + t_{\nu,0.975} \times SE$, onde $(t_{\nu,0.975})$ 97,5% da distribuição t com ν graus de liberdade, usado para construir um intervalo de confiança de 95%. Para cada intervalo $[LI_i, LS_i]$, é verificado se o verdadeiro valor do parâmetro θ_i está dentro desses intervalos. Para cada simulação i consideramos a seguinte função indicadora,

$$I = \begin{cases} 1 & \text{se LI} \leq \theta_i \leq LS \\ .0 & \text{caso contrário.} \end{cases}$$

A CE é calculada como a proporção de intervalos de confiança que incluem o verdadeiro valor:

$$CE = \frac{1}{N}\sum_{i=1}^{N} I_i.$$

Desta forma, a CE varia de 0 até 1, onde um valor próximo de 0.95, por exemplo, indica que aproximadamente 95% dos intervalos de confiança gerados incluem o verdadeiro valor do parâmetro.

A diferença entre o EP e o DP é que o EP é calculado diretamente a partir do modelo ajustado, enquanto o DP é obtido empiricamente. Essa distinção é importante para verificarmos a consistência dos nossos estimadores, uma vez que, se as estimativas forem muito distintas entre os métodos, poderemos estar diante de uma subestimação da variância.

Para cada modelo de imputação aplicado, foram gerados $K = 5$ conjuntos de dados imputados e combinados via regras de Rubin. Aplicamos 2000 simulações de MC com tamanhos amostrais de 500 e 1000.

As simulações para o padrão de ausência MAR em X_1 resultaram em aproximadamente 41,1% em média e para o padrão MCAR 37,6% de ausência em média. Para X_2, sob mecanismo MAR, temos aproximadamente 50% de perda de dados em média. Diante do padrão de ausência MAR MNAR para X_1 e X_2 a perda geral da base foi de 72,8% em

média e para o padrão MCAR MNAR 69,2% em média. A proporção de censura resultante foi de 31,1% em média e o tempo de censura considerado foi $S_C(T) = exp(-T^{1/2})$, para o cálculo de δ consideramos a expressão (2.1). Considere os resultados destacados em negrito nas tabelas como as variáveis que sofreram o mecanismo de perda. O código que gerou os resultados da simulação, bem como as figuras ilustrativas, estão apresentados no Apêndice (Seção 7).

4.1 VARIÁVEIS EXPLICATIVAS X_1 E X_2 CONTÍNUAS

4.1.1 CENÁRIO 1 - X_1 MAR

Diante da análise da Tabela 2, onde a variável X_1 é parcialmente observada sob um mecanismo MAR, e as variáveis X_2 (contínua) e Z (binária) são completamente observadas, foram comparados quatro métodos de imputação de dados e em seguida ajustado o modelo de Cox para cada um dos bancos de dados resultantes com intuito de estimar os parâmetros θ_1, θ_2 e θ_3.

Tabela 2 – Resultados para ausência em X_1 (MAR)

		Estimativa de θ_1				Estimativa de θ_2				Estimativa de ,θ_3			
n	Método	**Viés**	**EP**	**DP**	**CE**	Viés	EP	DP	CE	Viés	EP	DP	CE
500	1. FULL	**0,010**	**0,078**	**0,081**	**0,944**	0,006	0,078	0,081	0,940	0,009	0,134	0,136	0,944
	2. CC	**0,073**	**0,105**	**0,109**	**0,895**	0,071	0,105	0,109	0,898	0,070	0,178	0,181	0,925
	3. NORM	**-0,294**	**0,106**	**0,075**	**0,178**	-0,150	0,090	0,079	0,600	-0,149	0,154	0,135	0,849
	4. CONG	**0,018**	**0,101**	**0,111**	**0,928**	0,011	0,091	0,094	0,940	0,013	0,153	0,158	0,940
	5. NA	**-0,188**	**0,108**	**0,084**	**0,576**	-0,113	0,093	0,080	0,780	-0,112	0,158	0,139	0,910
	6. CART	**-0,068**	**0,094**	**0,103**	**0,852**	-0,063	0,086	0,091	0,866	-0,095	0,145	0,150	0,887
1000	1. FULL	**0,004**	**0,055**	**0,055**	**0,942**	0,004	0,055	0,055	0,951	0,006	0,094	0,095	0,946
	2. CC	**0,062**	**0,073**	**0,073**	**0,881**	0,061	0,073	0,072	0,884	0,065	0,124	0,126	0,908
	3. NORM	**-0,306**	**0,076**	**0,053**	**0,011**	-0,155	0,063	0,054	0,290	-0,156	0,109	0,097	0,722
	4. CONG	**0,001**	**0,071**	**0,082**	**0,927**	0,003	0,063	0,066	0,933	0,004	0,107	0,111	0,939
	5. NA	**-0,196**	**0,077**	**0,057**	**0,242**	-0,118	0,065	0,056	0,565	-0,118	0,111	0,100	0,826
	6. CART	**-0,056**	**0,066**	**0,074**	**0,814**	-0,049	0,061	0,065	0,841	-0,078	0,102	0,112	0,851

Para uma amostra de $n = 500$, a estimativa de θ_1 pelo método NORM apresentou o maior viés negativo (-0,294), seguido pelo método NA (-0,188). A taxa de cobertura para o método NORM foi a mais baixa entre os métodos analisados, sendo de apenas 17,8%.

Na estimativa de θ_2, o método NORM também apresentou um viés negativo considerável (-0,150), embora sua taxa de cobertura tenha sido maior (60,0%) em comparação com θ_1. Por outro lado, o método CONG apresentou um viés praticamente nulo (0,011) e taxas de cobertura próximas dos níveis nominais (94,0%). Para θ_3, o padrão se repete com o método NORM apresentando um viés negativo (-0,149), enquanto o método CONG novamente demonstra um viés próximo de zero (0,013) e uma taxa de cobertura de 94,0%.

Na segunda parte da tabela, com $n = 1000$, a estimativa de θ_1 pelo método NORM apresentou um viés negativo ainda maior (-0,306), com uma taxa de cobertura drastica-

mente reduzido (1,1%). O método CONG, por sua vez, manteve um desempenho estável, com viés muito próximo de zero (0,001) e uma taxa de cobertura elevada (92,7%).

Finalmente, para θ_2 e θ_3, os métodos NORM e NA continuaram a apresentar vieses negativos, enquanto o método CONG mostrou vieses mínimos e taxas de cobertura próximas de 93%, confirmando sua consistência e robustez frente às dificuldades impostas pelo mecanismo de ausência MAR que incluem a necessidade de fazer suposições não testáveis sobre o processo de dados ausentes. Além disso, estamos sob a suposição de que as probabilidades de não resposta não dependem de nenhuma informação não observada, o desafio é escolher um conjunto útil de preditores de imputação quando temos um grande conjunto de variáveis e a complexidade de gerar imputações reais quando temos diferentes classes de variáveis.

De acordo com a Figura 4, para o gráfico *(a)* o valor verdadeiro do parâmetro é representado pela linha constante tracejada no valor 1. Cada resultado é visualizado em um gráfico de violino, que ilustra a distribuição das estimativas médias dos parâmetros para cada método de tratamento de dados ausentes. Observa-se que, apesar do método NORM apresentar estimativas com maior viés, a distribuição dessas estimativas possui variações menores em comparação com o método CC e outros métodos de imputação. É importante notar que as estimativas para Z, variável completamente observada e binária, exibem maior variabilidade em comparação com a variável contínua X_2, que também é completamente observada neste cenário. O gráfico *(b)* mostra a taxa de cobertura para cada método de imputação em formato de barras, com uma linha tracejada fixada em 0,95. É evidente que o método CONG é o que mais se aproxima dos valores nominais de cobertura. Além disso, alguns modelos, como NORM e NA, apresentaram taxas de cobertura inferiores às do caso completo, indicando uma subestimação da incerteza associada com esses métodos.

4.1.2 CENÁRIO 2 - X_2 MAR

Considerando o cenário onde a variável X_1 é parcialmente observada sob um mecanismo MAR, e as variáveis X_2 (contínua) e Z (binária) são completamente observadas. Para uma amostra de $n = 500$, a estimativa de θ_1 pelo método CC apresentou o maior viés positivo (0,176), enquanto o método CONG demonstrou um viés praticamente nulo (0,006), aliado a uma taxa de cobertura (94,2%). O método NORM, embora tenha mostrado um viés negativo moderado (-0,074), ainda obteve uma boa taxa de cobertura (88,4%) (Tabela 3).

Tabela 3 – Resultados para ausência em X_2 (MAR)

		Estimativa de θ_1				**Estimativa de θ_2**				Estimativa de θ_3			
n	Método	Viés	EP	DP	CE	**Viés**	**EP**	**DP**	**CE**	Viés	EP	DP	CE
500	1. FULL	0,009	0,078	0,078	0,947	**0,007**	**0,078**	**0,078**	**0,947**	0,008	0,134	0,135	0,950
	2. CC	0,176	0,120	0,125	0,699	**0,063**	**0,107**	**0,108**	**0,923**	0,121	0,185	0,192	0,901
	3. NORM	-0,074	0,092	0,082	0,884	**-0,258**	**0,105**	**0,078**	**0,282**	-0,110	0,155	0,136	0,916
	4. CONG	0,006	0,095	0,096	0,942	**0,015**	**0,104**	**0,109**	**0,934**	0,011	0,154	0,157	0,945
	5, NA	-0,033	0,096	0,089	0,936	**-0,105**	**0,108**	**0,089**	**0,848**	-0,056	0,159	0,144	0,948
	6. CART	-0,067	0,082	0,095	0,826	**-0,037**	**0,093**	**0,119**	**0,845**	-0,084	0,143	0,155	0,877
1000	1. FULL	0,002	0,054	0,055	0,946	**0,004**	**0,055**	**0,055**	**0,947**	0,001	0,094	0,095	0,950
	2. CC	0,162	0,082	0,086	0,505	**0,055**	**0,074**	**0,078**	**0,886**	0,111	0,129	0,132	0,861
	3. NORM	-0,083	0,065	0,057	0,750	**-0,268**	**0,075**	**0,055**	**0,032**	-0,121	0,109	0,096	0,826
	4. CONG	-0,002	0,066	0,070	0,927	**-0,000**	**0,072**	**0,090**	**0,919**	-0,002	0,108	0,113	0,932
	5. NA	-0,040	0,067	0,063	0,911	**-0,111**	**0,076**	**0,062**	**0,697**	-0,066	0,112	0,102	0,921
	6. CART	-0,050	0,058	0,071	0,805	**-0,022**	**0,065**	**0,087**	**0,850**	-0,068	0,100	0,116	0,851

Na estimativa de θ_2, o método NORM apresentou o maior viés negativo (-0,258) e uma das menores taxas de cobertura (28,2%), indicando um desempenho inferior neste cenário. Em contrapartida, o método CONG novamente se destacou, com um viés mínimo (0,015) e uma taxa de cobertura de 93,4%, sugerindo um desempenho robusto. Para θ_3, o método CONG continuou a apresentar viés próximo de zero (0,011) e uma alta taxa de cobertura (94,5%), enquanto o método CART apresentou um viés negativo (-0,084) e uma taxa de cobertura inferior (87,7%).

Na segunda parte da tabela, com $n = 1000$, a estimativa de θ_1 pelo método CC manteve um viés positivo considerável (0,162) e uma taxa de cobertura significativamente baixa (50,5%). O método CONG, por outro lado, demonstrou um desempenho estável com viés próximo de zero (-0,002) e uma taxa de cobertura elevada (92,7%).

Finalmente, para θ_2 e θ_3, os métodos NORM e NA continuaram a apresentar vieses negativos, especialmente o NORM para θ_2 (-0,268) com uma taxa de cobertura drasticamente reduzida (3,2%). O método CONG, em contraste, mostrou vieses mínimos e taxas de cobertura próximas de 93%, confirmando sua consistência e robustez em lidar com o padrão de ausência MAR dependente de múltiplas covariáveis.

Podemos observar que a distribuição do método NORM é a que mais se diferencia das demais e oferece estimativas mais viesadas. É possível notar que no método CART em média as estimativas ficam mais próximas do verdadeiro valor do parâmetro. O mesmo ocorre para o método CONG. Conforme mencionado anteriormente os métodos NA e CART apresentaram menores taxas de cobertura quando comparado ao caso completo que possui como característica viés nas estimativas. Dentre os métodos de imputação, novamente, o CONG apresentou taxas de cobertura próximas dos níveis nominais (Figura 5).

4.1.3 CENÁRIO 3 - X_1 MCAR

Para uma amostra de $n = 500$, a estimativa de θ_1 pelo método NORM apresentou o maior viés negativo (-0,321), com uma taxa de cobertura de 13,1%, a menor entre os

métodos analisados. O método CONG, por outro lado, apresentou um viés muito baixo (0,012) e uma taxa de cobertura razoável (93,8%). Para θ_2, o método NORM também exibiu um viés negativo considerável (-0,170) e uma baixa taxa de cobertura (52,2%). O método CONG teve um viés menor (0,010) e uma taxa de cobertura de 93,9%, mostrando um desempenho superior. No caso de θ_3, o método CONG continuou a demonstrar um viés baixo (0,006) e uma taxa de cobertura de 94,9%, enquanto o método CART apresentou um viés negativo (-0,124) e uma taxa de cobertura de 85,1% (Tabela 4).

Tabela 4 – Resultados para ausência em X_1 (MCAR)

		Estimativa de θ_1				Estimativa de θ_2				Estimativa de θ_3			
n	Método	**Viés**	**EP**	**DP**	**CE**	Viés	EP	DP	CE	Viés	EP	DP	CE
500	1. FULL	**0,007**	**0,078**	**0,078**	**0,952**	0,008	0,078	0,081	0,942	0,005	0,134	0,132	0,948
	2. CC	**0,012**	**0,102**	**0,103**	**0,947**	0,013	0,103	0,105	0,948	0,007	0,175	0,175	0,950
	3. NORM	**-0,321**	**0,108**	**0,080**	**0,131**	-0,170	0,090	0,083	0,522	-0,173	0,158	0,132	0,839
	4. CONG	**0,012**	**0,101**	**0,107**	**0,938**	0,010	0,091	0,096	0,939	0,006	0,156	0,154	0,949
	5. NA	**-0,212**	**0,110**	**0,083**	**0,506**	-0,137	0,094	0,084	0,670	-0,142	0,162	0,132	0,888
	6. CART	**-0,096**	**0,096**	**0,102**	**0,793**	-0,073	0,087	0,093	0,826	-0,124	0,147	0,148	0,851
1000	1. FULL	**0,003**	**0,055**	**0,054**	**0,948**	0,002	0,054	0,055	0,952	0,003	0,094	0,095	0,951
	2. CC	**0,005**	**0,071**	**0,073**	**0,946**	0,004	0,071	0,071	0,948	0,008	0,122	0,123	0,951
	3. NORM	**-0,334**	**0,076**	**0,057**	**0,005**	-0,179	0,063	0,057	0,184	-0,179	0,111	0,098	0,650
	4. CONG	**-0,006**	**0,071**	**0,100**	**0,898**	-0,001	0,064	0,072	0,921	0,000	0,109	0,116	0,925
	5. NA	**-0,219**	**0,078**	**0,058**	**0,160**	-0,145	0,066	0,057	0,410	-0,146	0,115	0,100	0,768
	6. CART	**-0,075**	**0,067**	**0,074**	**0,754**	-0,060	0,061	0,065	0,796	-0,102	0,103	0,114	0,805

Na segunda parte da tabela, com $n = 1000$, a estimativa de θ_1 pelo método NORM manteve um viés negativo acentuado (-0,334) e uma taxa de cobertura significativamente baixa (0,5%). O método CONG apresentou um viés próximo de zero (-0,006) e uma taxa de cobertura de 89,8%, demonstrando melhor desempenho. Para θ_2, o método NORM exibiu um viés negativo (-0,179) e uma taxa de cobertura baixa (18,4%), enquanto o método CONG teve um viés muito baixo (-0,001) e uma taxa de cobertura de 92,1%. Para θ_3, o método CONG continuou a apresentar um viés mínimo e uma taxa de cobertura de 92,5%, enquanto o método CART mostrou um viés negativo (-0,102) e uma taxa de cobertura de 80,5% (Figura 6).

Todavia, neste cenário onde a ausência é do tipo MCAR a análise CC é consistente pois a natureza da ausência não está dependendo de nenhuma variável completamente ou parcialmente observada. Neste sentido, apesar do CONG se destacar entre os modelos de imputação, ainda assim não seria necessário utilizar o tratamento via IM. Estes resultados mostram que em casos onde a ausência é completamente aleatória a imputação pode piorar as estimativas e gerar taxas de coberturas abaixo dos níveis nominais.

4.1.4 CENÁRIO 4 - X_1 MAR E X_2 MNAR

Após a avaliação das perdas individuais, passamos a analisar a perda conjunta de X_1 e X_2 (Tabela 5). Como X_2 também apresenta dados faltantes e a ausência é dependente de uma variável parcialmente observada (X_1), estamos lidando com um padrão MNAR.

Nesta situação, não é possível separar os efeitos das ausências em X_1 e X_2, resultando em um efeito confundido de ausência para ambas as variáveis.

Tabela 5 – Resultados para ausência em X_1 e X_2 (MAR, MNAR)

		Estimativa de θ_1				**Estimativa de θ_2**				Estimativa de θ_3			
n	Método	**Viés**	**EP**	**DP**	**CE**	**Viés**	**EP**	**DP**	**CE**	Viés	EP	DP	CE
500	1. FULL	**0,005**	**0,078**	**0,076**	**0,959**	**0,006**	**0,078**	**0,078**	**0,949**	0,005	0,134	0,137	0,946
	2. CC	**0,270**	**0,162**	**0,167**	**0,623**	**0,127**	**0,146**	**0,148**	**0,883**	0,196	0,248	0,255	0,892
	3. NORM	**-0,323**	**0,110**	**0,082**	**0,150**	**-0,343**	**0,109**	**0,084**	**0,109**	-0,225	0,166	0,139	0,740
	4. CONG	**-0,024**	**0,117**	**0,119**	**0,924**	**0,005**	**0,117**	**0,123**	**0,923**	-0,003	0,170	0,180	0,934
	5. NA	**-0,206**	**0,117**	**0,097**	**0,555**	**-0,178**	**0,116**	**0,098**	**0,657**	-0,156	0,174	0,153	0.869
	6. CART	**-0,152**	**0,109**	**0,111**	**0,682**	**-0,103**	**0,112**	**0,119**	**0,797**	-0,176	0,159	0,156	0,785
1000	1. FULL	**0,003**	**0,055**	**0,054**	**0,951**	**0,000**	**0,054**	**0,055**	**0,951**	-0,003	0,094	0,096	0,941
	2. CC	**0,251**	**0,110**	**0,116**	**0,379**	**0,118**	**0,100**	**0,103**	**0,806**	0,190	0,171	0,174	0,811
	3. NORM	**-0,331**	**0,077**	**0,058**	**0,009**	**-0,352**	**0,076**	**0,058**	**0,004**	-0,234	0,117	0,093	0,467
	4. CONG	**-0,033**	**0,080**	**0,087**	**0,907**	**-0,004**	**0,081**	**0,089**	**0,930**	-0,011	0,120	0,125	0,927
	5. NA	**-0,211**	**0,082**	**0,068**	**0,258**	**-0,181**	**0,082**	**0,066**	**0,377**	-0,161	0,122	0,102	0,743
	6. CART	**-0,129**	**0,078**	**0,080**	**0,592**	**-0,081**	**0,081**	**0,084**	**0,784**	-0,151	0,115	0,115	0,707

Observa-se que, embora o método NORM mostre estimativas mais viciadas (Viés = $-0{,}323$ para θ_1 e $-0{,}343$ para θ_2 com 500 amostras), o viés dessas estimativas é menor comparada ao método CC e outros métodos de imputação. O método CC, por exemplo, apresenta viés de 0,270 para θ_1 e 0,127 para θ_2 com 500 amostras.

O método CONG se destaca por apresentar taxas de cobertura mais próximas dos valores nominais (92,4% para θ_1 e 92,3% para θ_2 com 500 amostras). Outros métodos, como NORM e NA, mostraram taxas de cobertura inferiores ao caso completo (Figura 7).

Para $n = 1000$, a análise CC apresentou viés, principalmente em θ_1 (0,183). Entre os métodos de imputação, o método NORM mostrou-se pior que o CC, com vieses de -0,350 e -0,359 para θ_1 e θ_2, respectivamente. O NA apresentou o segundo pior viés, tanto para θ_1 quanto para θ_2. Por outro lado, o CART apresentou menor viés que o CC. Em relação às taxas de cobertura, o pior desempenho foi observado no modelo de imputação NORM, seguido pelo método NA, e depois pelo CART. Nesse contexto, a análise CC apresentou melhores taxas de cobertura. Mais uma vez, o modelo CONG obteve estimativas menos viesadas e taxas de cobertura próximas dos níveis nominais.

4.1.5 CENÁRIO 5 - X_1 MCAR E X_2 MNAR

Para verificar se a mudança no padrão de ausência impactaria as conclusões anteriores, consideramos o caso em que X_1 está ausente completamente ao acaso (MCAR), enquanto X_2 continua a apresentar padrão MNAR, dado que depende de uma covariável parcialmente observada (Tabela 6).

Tabela 6 – Resultados para ausência em X_1 e X_2 (MCAR, MNAR)

		Estimativa de θ_1				**Estimativa de θ_2**				Estimativa de θ_3			
n	Método	**Viés**	**EP**	**DP**	**CE**	**Viés**	**EP**	**DP**	**CE**	Viés	EP	DP	CE
500	1. FULL	**0,008**	**0,078**	**0,078**	**0,951**	**0,008**	**0,078**	**0,077**	**0,952**	0,014	0,134	0,136	0,941
	2. CC	**0,183**	**0,159**	**0,166**	**0,821**	**0,077**	**0,143**	**0,149**	**0,921**	0,130	0,245	0,253	0,921
	3. NORM	**-0,350**	**0,111**	**0,085**	**0,104**	**-0,359**	**0,110**	**0,084**	**0,083**	-0,244	0,170	0,133	0,726
	4. CONG	**-0,026**	**0,115**	**0,120**	**0,916**	**0,020**	**0,117**	**0,127**	**0,922**	0,004	0,174	0,180	0,939
	5. NA	**-0,223**	**0,118**	**0,094**	**0,502**	**-0,187**	**0,117**	**0,096**	**0,641**	-0,173	0,177	0,148	0,862
	6. CART	**-0,179**	**0,111**	**0,107**	**0,605**	**-0,104**	**0,114**	**0,115**	**0,811**	-0,192	0,164	0,153	0,779
1000	1. FULL	**0,001**	**0,054**	**0,056**	**0,941**	**0,002**	**0,055**	**0,055**	**0,954**	0,004	0,094	0,093	0,954
	2. CC	**0,167**	**0,109**	**0,113**	**0,678**	**0,057**	**0,097**	**0,099**	**0,916**	0,113	0,168	0,165	0,910
	3. NORM	**-0,363**	**0,079**	**0,062**	**0,004**	**-0,373**	**0,078**	**0,061**	**0,002**	-0,254	0,119	0,094	0,403
	4. CONG	**-0,041**	**0,080**	**0,090**	**0,890**	**-0,002**	**0,082**	**0,097**	**0,910**	-0,011	0,120	0,123	0,938
	5, NA	**-0,234**	**0,084**	**0,068**	**0,197**	**-0,199**	**0,083**	**0,069**	**0,321**	-0,183	0,125	0,101	0,722
	6. CART	**-0,151**	**0,079**	**0,081**	**0,514**	**-0,083**	**0,081**	**0,088**	**0,773**	-0,167	0,118	0,117	0,693

No cenário com padrão MCAR para X_1, a análise do método CC indica uma redução no viés das estimativas e taxas de cobertura mais próximas aos níveis nominais em comparação com o padrão MAR. O viés para θ_1 no método CC reduz para 0,183 com 500 amostras e a taxa de cobertura sobe para 82,1%. O método NORM mostra estimativas e taxas de cobertura semelhantes, sem melhora ou piora com a mudança para o padrão MCAR de X_1 (viés de $-0{,}350$ para θ_1 com 500 amostras).

Os métodos NA, CART e CONG apresentam estimativas ligeiramente inferiores com a ausência de X_1 como MCAR. No entanto, o método CONG continua a fornecer estimativas menos viesadas e taxas de cobertura próximas dos níveis nominais (91,6% para θ_1 e 92,2% para θ_2 com 500 amostras) (Figura 8).

4.2 VARIÁVEIS EXPLICATIVAS X_1 E X_2 BINÁRIAS

4.2.1 CENÁRIO 1 - X_1 MAR

Considerando uma perda MAR para X_1 binário, dependente do tempo, e levando em conta que X_2 e Z são completamente observados, analisamos os resultados dos diferentes métodos de imputação.

Os resultados mostram que o método LOG foi o que apresentou o maior viés na estimativa média dos parâmetros, com um viés de -0,121 para θ_1 e taxas de cobertura de 88,3% para a amostra de 500. Para θ_2, o viés foi de -0,048 e a taxa de cobertura foi de 94,7%. Já para θ_3, o viés foi de -0,053 com uma cobertura de 93,9% (Tabela 7).

Tabela 7 – Resultados para ausência em X_1

		Estimativa de θ_1				Estimativa de θ_2				Estimativa de θ_3			
n	Método	**Viés**	**EP**	**DP**	**CE**	Viés	EP	DP	CE	Viés	EP	DP	CE
500	1. FULL	**0,006**	**0,120**	**0,121**	**0,949**	0,012	0,120	0,120	0,949	0,007	0,120	0,121	0,957
	2. CC	**0,068**	**0,164**	**0,164**	**0,935**	0,073	0,164	0,167	0,927	0,067	0,164	0,168	0,932
	3. LOG	**-0,121**	**0,159**	**0,150**	**0,883**	-0,048	0,128	0,121	0,947	-0,053	0,128	0,122	0,939
	4. CONG	**0,016**	**0,155**	**0,156**	**0,944**	0,014	0,130	0,130	0,963	0,008	0,130	0,133	0,948
	5. NA	**-0,078**	**0,159**	**0,151**	**0,920**	-0,040	0,129	0,122	0,945	-0,045	0,129	0,124	0,938
	6. CART	**-0,044**	**0,140**	**0,166**	**0,885**	-0,021	0,125	0,132	0,933	-0,028	0,125	0,135	0,926
1000	1. FULL	**0,004**	**0,084**	**0,082**	**0,955**	0,006	0,084	0,083	0,952	0,004	0,084	0,083	0,950
	2. CC	**0,066**	**0,115**	**0,112**	**0,915**	0,065	0,115	0,115	0,915	0,062	0,115	0,114	0,922
	3. LOG	**-0,125**	**0,111**	**0,102**	**0,785**	-0,054	0,090	0,083	0,918	-0,055	0,090	0,084	0,922
	4. CONG	**0,010**	**0,108**	**0,106**	**0,944**	0,008	0,092	0,091	0,951	0,006	0,091	0,091	0,944
	5. NA	**-0,085**	**0,111**	**0,102**	**0,891**	-0,047	0,091	0,084	0,932	-0,047	0,091	0,085	0,929
	6. CART	**-0,030**	**0,098**	**0,114**	**0,886**	-0,016	0,088	0,094	0,925	-0,019	0,088	0,094	0,924

A análise CC, por sua vez, teve desempenhos melhores em relação ao LOG, com vieses de 0,068 para θ_1, 0,073 para θ_2, e 0,067 para θ_3 nas amostras de 500, com taxas de cobertura de 93,5%, 92,7% e 93,2%, respectivamente. Comparado ao LOG, o CC forneceu estimativas mais próximas dos valores verdadeiros e com taxas de cobertura mais confiáveis.

Os métodos NA e CART apresentaram estimativas menos viesadas em comparação ao LOG, mas o CART teve taxas de cobertura inferiores. O método NA mostrou vieses de -0,078 para θ_1, -0,040 para θ_2, e -0,045 para θ_3 com taxas de cobertura de 92,0%, 94,5% e 93,8%, respectivamente. O método CART teve vieses de -0,044 para θ_1, -0,021 para θ_2, e -0,028 para θ_3, com taxas de cobertura de 88,5%, 93,3% e 92,6%. Assim, o método NA teve taxas de cobertura relativamente melhores comparadas ao CART, especialmente para θ_1 e θ_2.

Para $n = 1000$ a distribuição das estimativas para o método usual (LOG) apresentou menor viés quando comparada aos dados contínuos (NORM), embora, para X_1 binário, os resultados foram mais favoráveis com taxas de cobertura de 78,5% para θ_1, 91,8% para θ_2, e 92,2% para θ_3. O método CONG se destacou, apresentando menores vieses e taxas de cobertura próximas dos níveis aceitáveis, com coberturas de 94,4% para θ_1, 95,1% para θ_2, e 94,4% para θ_3 (Figura 9).

4.2.2 CENÁRIO 2 - X_2 MAR

Quando consideramos a ausência de X_2 e a completa observação de X_1 e Z, a análise revela que a análise CC apresenta o maior viés nas estimativas. O método LOG, juntamente com o CART, demonstrou estimativas menos viesadas comparado aos métodos CONG e NA. No geral, as taxas de cobertura foram próximas entre os métodos, mas o CART teve a taxa de cobertura mais baixa, em torno de 90% (Tabela 8).

Tabela 8 – Estimativa média dos parâmetros X_2 (MAR)

		Estimativa de θ_1				**Estimativa de θ_2**				Estimativa de θ_3			
n	Método	Viés	EP	DP	CE	**Viés**	**EP**	**DP**	**CE**	Viés	EP	DP	CE
500	1. FULL	-0,000	0,119	0,123	0,946	**0,009**	**0,120**	**0,122**	**0,947**	0,007	0,120	0,121	0,943
	2. CC	0,126	0,161	0,161	0,896	**0,052**	**0,148**	**0,150**	**0,934**	0,084	0,152	0,155	0,923
	3. LOG	-0,026	0,126	0,126	0,945	**0,005**	**0,147**	**0,152**	**0,939**	-0,029	0,126	0,122	0,952
	4. CONG	-0,001	0,127	0,131	0,944	**0,017**	**0,143**	**0,144**	**0,944**	0,007	0,126	0,128	0,944
	5. NA	-0,026	0,127	0,128	0,942	**0,021**	**0,147**	**0,148**	**0,940**	-0,025	0,126	0,122	0,949
	6. CART	-0,025	0,123	0,133	0,922	**-0,005**	**0,135**	**0,155**	**0,904**	-0,016	0,123	0,129	0,936
1000	1. FULL	0,006	0,084	0,086	0,940	**0,005**	**0,084**	**0,086**	**0,947**	0,002	0,084	0,084	0,953
	2. CC	0,129	0,113	0,114	0,805	**0,047**	**0,104**	**0,104**	**0,925**	0,075	0,106	0,109	0,894
	3. LOG	-0,020	0,089	0,089	0,937	**-0,001**	**0,103**	**0,106**	**0,939**	-0,033	0,088	0,086	0,938
	4. CONG	0,005	0,089	0,091	0,938	**0,011**	**0,101**	**0,103**	**0,942**	0,002	0,089	0,090	0,946
	5. NA	-0,021	0,089	0,089	0,935	**0,015**	**0,103**	**0,104**	**0,945**	-0,029	0,089	0,086	0,943
	6. CART	-0,008	0,086	0,094	0,922	**-0,005**	**0,094**	**0,108**	**0,913**	-0,012	0,086	0,092	0,933

Para a amostra de 500, o método CC apresentou vieses mais elevados: 0,126 para θ_1, 0,052 para θ_2, e 0,084 para θ_3, com taxas de cobertura de 89,6%, 93,4%, e 92,3%.

Os métodos LOG e CART mostraram vieses relativamente menores. O método LOG teve vieses de -0,026 para θ_1, 0,005 para θ_2, e -0,029 para θ_3, com taxas de cobertura de 94,5%, 93,9%, e 95,2%, respectivamente. O método CART apresentou vieses de -0,025 para θ_1, -0,005 para θ_2, e -0,016 para θ_3, com taxas de cobertura de 92,2%, 90,4%, e 93,6%. Esses resultados destacam que, embora o LOG e o CART apresentem estimativas mais precisas, o CART tende a ter taxas de cobertura mais baixas. Isso se deve ao fato de que o método de imputação via CART pode não recriar com precisão a distribuição condicional de determinadas variáveis, levando a uma especificação incorreta do modelo de imputação para essas variáveis.

O método CONG teve desempenho consistente com vieses de -0,001 para θ_1, 0,017 para θ_2, e 0,007 para θ_3, e taxas de cobertura de 94,4%, 94,4%, e 94,4%. O método NA também apresentou vieses relativamente baixos: -0,026 para θ_1, 0,021 para θ_2, e -0,025 para θ_3, com taxas de cobertura de 94,2%, 94,0%, e 94,9%.

Para $n = 1000$ a análise CC apresenta estimativas mais viesadas para θ_2 (0,047) e taxas de cobertura razoáveis (92,5%). Todavia, as melhores taxas de cobertura podem ser observadas nos modelos de imputação NA (95,5%) e CONG (94,2%). As estimativas menos viesadas para θ_2 é no modelo LOG (-0,001) e CONG (0,011).

Comparando com o mecanismo de ausência MAR dependente apenas do tempo, os métodos LOG e CART mostraram uma piora nas taxas de cobertura e estimativas menos precisas. O método CONG, em contraste, teve desempenho semelhante em ambos os cenários de ausência. Os métodos LOG e NA, por outro lado, mostraram uma média de estimativas mais viesadas e taxas de cobertura menores quando comparados ao mecanismo MAR dependente do tempo (Figura 10).

4.2.3 CENÁRIO 3 - X_1 MCAR

Quando X_1 está ausente de maneira MCAR e X_2 e Z estão completamente observados, a análise revela que a análise CC e LOG têm os maiores vieses entre os tratamentos analisados. A análise CC apresenta um viés de 0,212 para θ_1, 0,109 para θ_2, e 0,153 para θ_3, com taxas de cobertura de 85,9%, 91,7%, e 90,6%, respectivamente. Em contraste, o método LOG mostrou vieses de -0,212 para θ_1, -0,154 para θ_2, e -0,095 para θ_3, com taxas de cobertura de 75,0%, 85,3%, e 90,7%. .

O método NA teve uma taxa de cobertura melhor que a análise de casos completos, com vieses de -0,127 para θ_1, -0,070 para θ_2, e -0,076 para θ_3, e taxas de cobertura de 88,3%, 93,5%, e 92,6%. As taxas de cobertura para NA são próximas das do método CART, que apresentou vieses de -0,067 para θ_1, -0,034 para θ_2, e -0,052 para θ_3, e taxas de cobertura de 87,1%, 90,9%, e 91,4%. Embora o CART tenha uma taxa de cobertura mais baixa, seus vieses são menores que o do método CONG, que apresentou vieses de 0,009 para θ_1, 0,021 para θ_2, e 0,003 para θ_3, com taxas de cobertura de 93,1%, 93,1%, e 95,2%.

Ao comparar os cenários MCAR e MAR para a ausência de X_1, observamos que o caso CC tem estimativas mais viesadas no cenário MAR, com viés de 0,212 comparado a 0,068 no cenário MAR. Esse aumento no viés é acompanhado por um aumento na métrica SE, isto é, há uma maior incerteza na estimativa do parâmetro resultando em uma baixa nas taxas de cobertura. O mesmo padrão é observado para os métodos LOG, NA e CART. O método CONG, por outro lado, apresenta vieses menores no cenário MCAR, mas com uma pequena redução na taxa de cobertura. Está comparação faz sentido pois temos as mesmas covariáveis completamente observadas sob controle e as ausências ocorrem em apenas uma variável explicativa por vez, ou seja, podemos comparar os mecanismos neste contexto específico.

Comparando a ausência MCAR com variáveis contínuas e binárias, para a análise CC, as estimativas são mais viesadas para variáveis binárias e as taxas de cobertura se aproximam mais dos níveis nominais. Para os modelos de imputação NORM e LOG, o viés é menor para variáveis binárias, embora as taxas de cobertura ainda não estejam nos níveis nominais, mostrando uma melhoria conforme a natureza da variável imputada. No método CONG, as estimativas permanecem próximas e as taxas de cobertura são similares para variáveis de classes distintas. O método NA segue um padrão semelhante ao NORM e LOG. O método CART apresenta menor viés para variáveis contínuas, mas com taxas de cobertura inferiores quando comparado a variáveis binárias (Figura 11).

Tabela 9 – Resultados para ausência em X_1 (MCAR)

		Estimativa de θ_1				Estimativa de θ_2				Estimativa de θ_3			
n	Método	**Viés**	**EP**	**DP**	**CE**	Viés	EP	DP	CE	Viés	EP	DP	CE
500	1. FULL	**0,009**	**0,120**	**0,119**	**0,954**	0,005	0,120	0,121	0,948	0,007	0,120	0,116	0,950
	2. CC	**0,013**	**0,156**	**0,158**	**0,953**	0,007	0,156	0,159	0,954	0,013	0,156	0,159	0,944
	3. LOG	**-0,184**	**0,157**	**0,130**	**0,806**	-0,066	0,128	0,120	0,928	-0,060	0,128	0,118	0,934
	4. CONG	**0,015**	**0,155**	**0,160**	**0,942**	0,004	0,130	0,130	0,953	0,010	0,130	0,129	0,949
	5. NA	**-0,112**	**0,160**	**0,137**	**0,902**	-0,054	0,129	0,122	0,937	-0,049	0,129	0,120	0,939
	6. CART	**-0,048**	**0,141**	**0,164**	**0,888**	-0,029	0,125	0,132	0,930	-0,025	0,125	0,131	0,932
1000	1. FULL	**0,005**	**0,084**	**0,086**	**0,945**	0,006	0,084	0,083	0,955	0,005	0,084	0,083	0,956
	2. CC	**0,006**	**0,109**	**0,113**	**0,949**	0,009	0,109	0,107	0,954	0,008	0,109	0,111	0,950
	3. LOG	**-0,191**	**0,110**	**0,094**	**0,589**	-0,064	0,090	0,082	0,896	-0,064	0,090	0,083	0,909
	4. CONG	**-0,005**	**0,109**	**0,126**	**0,906**	0,003	0,091	0,089	0,959	0,003	0,091	0,092	0,947
	5. NA	**-0,118**	**0,113**	**0,098**	**0,830**	-0,053	0,091	0,083	0,921	-0,052	0,091	0,085	0,929
	6. CART	**-0,036**	**0,098**	**0,117**	**0,882**	-0,017	0,088	0,092	0,936	-0,017	0,088	0,094	0,924

4.2.4 CENÁRIO 4 - X_1 MAR E X_2 MNAR

Observa-se que para $n = 500$ a estimativa de θ_1, o método LOG apresentou o maior viés negativo (-0,184), seguido pelo método NA (-0,112). A taxa de cobertura para o método LOG foi de 80,6% (Tabela 10).

Tabela 10 – Resultados para ausência em X_1 e X_2 (MAR, MNAR)

		Estimativa de θ_1				**Estimativa de θ_2**				Estimativa de θ_3			
n	Método	**Viés**	**EP**	**DP**	**CE**	**Viés**	**EP**	**DP**	**CE**	Viés	EP	DP	CE
500	1. FULL	**0,009**	**0,120**	**0,119**	**0,954**	**0,005**	**0,120**	**0,121**	**0,948**	0,007	0,120	0,116	0,950
	2. CC	**0,013**	**0,156**	**0,158**	**0,953**	**0,007**	**0,156**	**0,159**	**0,954**	0,013	0,156	0,159	0,944
	3. LOG	**-0,184**	**0,157**	**0,130**	**0,806**	**-0,066**	**0,128**	**0,120**	**0,928**	-0,060	0,128	0,118	0,934
	4. CONG	**0,015**	**0,155**	**0,160**	**0,942**	**0,004**	**0,130**	**0,130**	**0,953**	0,010	0,130	0,129	0,949
	5. NA	**-0,112**	**0,160**	**0,137**	**0,902**	**-0,054**	**0,129**	**0,122**	**0,937**	-0,049	0,129	0,120	0,939
	6. CART	**-0,048**	**0,141**	**0,164**	**0,888**	**-0,029**	**0,125**	**0,132**	**0,930**	-0,025	0,125	0,131	0,932
1000	1. FULL	**0,005**	**0,084**	**0,086**	**0,945**	**0,006**	**0,084**	**0,083**	**0,955**	0,005	0,084	0,083	0,956
	2. CC	**0,006**	**0,109**	**0,113**	**0,949**	**0,009**	**0,109**	**0,107**	**0,954**	0,008	0,109	0,111	0,950
	3. LOG	**-0,191**	**0,110**	**0,094**	**0,589**	**-0,064**	**0,090**	**0,082**	**0,896**	-0,064	0,090	0,083	0,909
	4. CONG	**-0,005**	**0,109**	**0,126**	**0,906**	**0,003**	**0,091**	**0,089**	**0,959**	0,003	0,091	0,092	0,947
	5. NA	**-0,118**	**0,113**	**0,098**	**0,830**	**-0,053**	**0,091**	**0,083**	**0,921**	-0,052	0,091	0,085	0,929
	6. CART	**-0,036**	**0,098**	**0,117**	**0,882**	**-0,017**	**0,088**	**0,092**	**0,936**	-0,017	0,088	0,094	0,924

Na estimativa de θ_2, o método LOG também apresentou um viés negativo (-0,066), embora a sua taxa de cobertura tenha sido maior (92,8%) em comparação com θ_1. O método CONG, por outro lado, mostrou um viés praticamente nulo (0,004) e a maior taxa de cobertura (95,3%), sugerindo um desempenho superior. Para θ_3, o padrão se repete com o método LOG apresentando viés negativo (-0,060), enquanto o método CONG novamente demonstra um viés próximo de zero (0,010) e uma alta taxa de cobertura (94,9%).

Na segunda parte da tabela, com $n = 1000$, a estimativa de θ_1 pelo método LOG apresentou um viés negativo ainda maior (-0,191), com uma taxa de cobertura drasticamente reduzida (58,9%). O método CONG, novamente, manteve um desempenho estável, com viés muito próximo de zero (-0,005) e uma taxa de cobertura razoável (90,6%).

Finalmente, para θ_2 e θ_3, os métodos LOG e NA continuaram a apresentar vieses negativos, enquanto o método CONG mostrou vieses mínimos e taxas de cobertura próximas de 95%, confirmando sua consistência e robustez frente às dificuldades impostas pelo mecanismo de faltas MAR (Figura 12).

4.2.5 CENÁRIO 5 - X_1 MCAR E X_2 MNAR

Quando $n = 500$, a análise CC apresentou um viés positivo significativo (0,212) e a menor taxa de cobertura (85,9%), indicando que a exclusão de casos pode introduzir viés considerável sob MCAR. Métodos como LOG e NA apresentaram vieses negativos (-0,212 e -0,127, respectivamente), com taxas de cobertura moderadas (75,0% e 88,3%).

Para θ_2 a análise CC, embora menos enviesado do que para θ_1, mostrou um viés positivo (0,109) e uma cobertura de 91,7%. O método LOG apresentou um viés negativo (-0,154) com uma cobertura de 85,3%, sugerindo uma ligeira subestimação.

Quanto à estimativa de θ_3, o tratamento CC apresentou um viés de 0,153 e uma cobertura de 90,6%, enquanto o método LOG teve um viés negativo (-0,095) e uma cobertura de 90,7%.

Quando se aumenta o tamanho da amostra para $n = 1000$, observa-se que o viés e a SD tendem a diminuir, isto indica menor variabilidade nas estimativas, sugerindo que o modelo está fornecendo resultados mais consistentes. A análise CC apresenta uma redução no viés, mas ainda assim é o método mais enviesado. O método LOG, apesar de apresentar uma melhoria na SD, ainda mostra vieses negativos significativos, especialmente para θ_3 (-0,096) (Tabela 11).

Tabela 11 – Resultados para ausência em X_1 e X_2 (MCAR, MNAR)

		Estimativa de θ_1				**Estimativa de θ_2**				Estimativa de θ_3			
n	Método	**Viés**	**EP**	**DP**	**CE**	**Viés**	**EP**	**DP**	**CE**	Viés	EP	DP	CE
500	1. FULL	**0,003**	**0,120**	**0,121**	**0,949**	**0,004**	**0,120**	**0,122**	**0,945**	0,004	0,120	0,121	0,953
	2. CC	**0,212**	**0,222**	**0,228**	**0,859**	**0,109**	**0,206**	**0,212**	**0,917**	0,153	0,210	0,209	0,906
	3. LOG	**-0,212**	**0,162**	**0,133**	**0,750**	**-0,154**	**0,155**	**0,133**	**0,853**	-0,095	0,133	0,121	0,907
	4. CONG	**0,009**	**0,165**	**0,170**	**0,931**	**0,021**	**0,159**	**0,167**	**0,931**	0,003	0,137	0,138	0,952
	5. NA	**-0,127**	**0,165**	**0,148**	**0,883**	**-0,070**	**0,157**	**0,142**	**0,935**	-0,076	0,135	0,124	0,926
	6. CART	**-0,067**	**0,149**	**0,172**	**0,871**	**-0,034**	**0,146**	**0,164**	**0,909**	-0,052	0,130	0,138	0,914
1000	1. FULL	**0,003**	**0,084**	**0,084**	**0,950**	**-0,000**	**0,084**	**0,084**	**0,958**	0,006	0,084	0,085	0,947
	2. CC	**0,201**	**0,154**	**0,153**	**0,764**	**0,099**	**0,143**	**0,144**	**0,900**	0,147	0,146	0,145	0,841
	3. LOG	**-0,220**	**0,114**	**0,090**	**0,523**	**-0,162**	**0,109**	**0,091**	**0,686**	-0,096	0,093	0,084	0,836
	4. CONG	**-0,002**	**0,114**	**0,121**	**0,933**	**0,011**	**0,111**	**0,118**	**0,920**	0,002	0,096	0,097	0,944
	5. NA	**-0,130**	**0,116**	**0,099**	**0,809**	**-0,074**	**0,111**	**0,098**	**0,906**	-0,075	0,095	0,086	0,884
	6. CART	**-0,048**	**0,105**	**0,118**	**0,879**	**-0,025**	**0,103**	**0,117**	**0,898**	-0,031	0,092	0,099	0,915

De maneira geral, os resultados mostraram que o desempenho dos métodos de imputação variou significativamente dependendo do mecanismo de dados ausentes e do tipo de variável. O método CONG geralmente teve um bom desempenho, produzindo estimativas menos viesadas e taxas de cobertura próxima aos níveis nominais pré-estabelecidos.

Em contraste, os métodos de regressão linear frequentemente resultaram em estimativas tendenciosas e baixas taxas de cobertura, particularmente em cenários MNAR.

5 APLICAÇÃO

Os dados empregados nesta pesquisa consistem em informações longitudinais de pacientes diagnosticados com doença de Chagas. A base de dados é composta por 619 observações e 28 variáveis. Se optássemos por excluir as linhas com dados ausentes da base, ou seja, se a escolha fosse pelo caso completo restariam 49 observações para a análise. A análise CC implica em aproximadamente 92% de perda de dados.

Selecionamos algumas variáveis do base de dados que apresentam ausência de dados para ilustrar o problema em estudos reais, tais como **Classe Funcional**, **Sorologia Chagas**, **Fração de Ejeção** e **Razão TEI**. Além disso, buscamos observar o impacto da imputação de dados em uma variável com mais de 50% de ausência, como a **Área do Ventrículo Direito em Sístole** (**AreaVDsis**), que possui aproximadamente 80% de valores ausentes. A tabela 12 mostra as variáveis selecionadas,

Tabela 12 – Descrição das Variáveis para Aplicação Prática

Variável	Descrição
Tempo	Tempo de vida em dias até a ocorrência do óbito do paciente.
Óbito	Morte do paciente (0 = Não Óbito, 1 = Óbito).
Classe Funcional	Classificação de insuficiência cardíaca da *New York Heart Association* (NYHA): 1 = Assintomático 2 = Levemente Sintomático 3 = Sintomático 4 = Sintomático em Repouso.
Sorologia Chagas	Classificação da doença de Chagas: 1 = Idiopático 2 = Chagas.
Fração de Ejeção	Percentual de sangue que o ventrículo esquerdo ejeta para a aorta durante a sístole.
Razão TEI	Índice de performance miocárdica que avalia a função sistólica do ventrículo direito.
Área VD em Sístole	Área do ventrículo direito em sístole.

O tempo zero é definido a partir do momento que o indivíduo procura tratamento no hospital. Conforme mencionado anteriormente, a variável que apresenta maior percentual de ausência é **AreaVDsis** a qual sua ausência ultrapassa 80% (Figura 3).

5.1 RECURSOS COMPUTACIONAIS

Para tratamento dos dados, bem como gráficos e tabelas utilizamos o pacote **tidyverse**. Os resultados das imputações serão obtidos utilizando o pacote **mice** o qual tem a implementação da imputação via regressão linear (NORM e LOG) e CART. O pacote **smcfcs** faz a implementação do modelo de imputação compatível com o modelo de análise (CONG). Para o ajuste do modelo de Cox o pacote **survival** foi aplicado. Todos os pacotes estão disponíveis no *software* R.

5.2 DADOS DE DOENÇA DE CHAGAS

O conjunto de dados aplicado na análise prática foi previamente tratado e as variáveis foram selecionadas conforme a descrição na página 45. Para compreender a probabilidade de sobrevivência ao longo do tempo foram construídas curvas de sobrevivência utilizando o estimador de Kaplan-Meier (KM). Este processo é realizado antes da aplicação de qualquer tratamento para os dados ausentes na base. As variáveis de natureza contínua **Fração de Ejeção**, **Razão TEI** e **AreaVDsis** foram categorizadas em intervalos.

A variável **Classe Funcional**, os indivíduos assintomáticos apresentam melhor curva de sobrevivência e os sintomáticos em repouso a pior. Em outras palavras, indivíduos assintomáticos têm uma probabilidade de sobrevivência mais alta em comparação com as outras classes ao longo do tempo, ou seja, indivíduos assintomáticos tendem a sobreviver por mais tempo antes de ocorrer o óbito. Já os indivíduos sintomáticos em repouso têm uma probabilidade de sobrevivência mais baixa em comparação às outras classes. Isso pode sugerir que eles venham a óbito mais rapidamente.

Para a **Razão TEI** a melhor curva se sobrevivência é no intervalo $< 1{,}7$, enquanto que a pior é em $> 2{,}3$. Em relação a **Fração de Ejeção** a pior curva de sobrevivência é em < 30 e a melhor em > 51. Para **Sorologia Chagas** as curvas ficam próximas em boa parte do tempo. Entretanto, a partir de $t = 4000$ notamos que o indivíduo idiopático apresenta valores superiores aos indivíduos chagásicos (Figura 2).

Cabe ressaltar que o eixo do tempo para as variáveis explicativas com ausência de dados, **Razão TEI** e **AreaVDsis**, ficam menores em comparação às variáveis explicativas completamente observadas. Neste sentido, podemos pensar que os últimos tempos estão associados aos valores não observados das variáveis explicativas.

A Tabela 13 apresenta uma análise detalhada dos percentuais de ausência e observação por categoria, focando em duas variáveis com ausência: **Razão TEI** e Área do Ventrículo Direito em Sístole (**AreaVDsis**). Para a **Razão TEI**, observamos variações entre os grupos analisados. No contexto de *Óbito*, a categoria *Não óbito* representa 36,00% das ausências e 64,00% das observações, enquanto a categoria *Óbito* inclui 17,67% das ausências e 82,33% das observações. Esses percentuais indicam uma predominância maior de observações na categoria de *Óbito* e um maior percentual de ausências na categoria de *Não óbito*.

Tabela 13 – Percentuais de ausência e observação por categoria de acordo com as variáveis ausentes

Razão TEI			
Categoria	**Ausências (Percentual)**	**Observações (Percentual)**	**Total**
Tempo (em dias)	Min = 0	Min = 0	-
	Max = 6857	Max = 6997	-
Óbito			
0 - Não óbito	135 (36,00%)	240 (64,00%)	375 (100%)
1 - Óbito	41 (17,67%)	191 (82,33%)	232 (100%)
Classe funcional			
1 - Assintomático	39 (20,87%)	148 (79,14%)	187 (100%)
2 - Levemente assintomático	99 (37,08%)	168 (62,92%)	267 (100%)
3 - Sintomático	27 (28,13%)	69 (71,88%)	96 (100%)
4 - Sintomático em repouso	11 (19,30%)	46 (80,70%)	57 (100%)
Sorologia Chagas			
1 - Idiopático	39 (26,90%)	106 (73,10%)	145 (100%)
2 - Chagas	137 (29,65%)	325 (70,35%)	462 (100%)
Fração de Ejeção	Min = 11,00	Min = 9,00	-
	Média = 35,00	Média = 43,00	-
	Max = 71,00	Max = 69,00	-
Área do ventrículo direito em sístole (AreaVDsis)			
Categoria	**Ausências (Percentual)**	**Observações (Percentual)**	**Total**
Tempo (em dias)	Min = 0	Min = 0	-
	Max = 3584	Max = 6997	-
Óbito			
0 - Não óbito	301 (80,00%)	74 (19,73%)	375 (100%)
1 - Óbito	216 (93,10%)	16 (6,90%)	232 (100%)
Classe funcional			
1 - Assintomático	154 (82,35%)	33 (17,65%)	187 (100%)
2 - Levemente assintomático	228 (85,39%)	39 (14,61%)	267 (100%)
3 - Sintomático	82 (85,42%)	14 (14,58%)	96 (100%)
4 - Sintomático em repouso	53 (92,98%)	4 (7,02%)	57 (100%)
Sorologia Chagas			
1 - Idiopático	127 (87,59%)	18 (12,41%)	145 (100%)
2 - Chagas	390 (84,42%)	72 (15,58%)	462 (100%)
Fração de Ejeção	Min = 18,00	Min = 9,00	-
	Média = 37,04	Média = 37,63	-
	Max = 60,00	Max = 71,00	-

Em relação à **Classe Funcional**, as diferenças entre os grupos também são notáveis. Considerando a ausência na variável **Razão TEI**, a categoria *Assintomático* (Classe 1) conta com 20,87% das ausências e 79,14% das observações. Já a categoria *Levemente assintomático*(Classe 2) apresenta 37,08% das ausências e 62,92% das observações. As categorias *Sintomático* (Classe 3) e *Sintomático em repouso* (Classe 4) mostram variações de 28,13% e 19,30% nas ausências, respectivamente, e 71,88% e 80,70% nas observações.

A **Sorologia Chagas** apresenta categorias distintas, com 26,90% das ausências e 73,10% das observações na classe *Idiopático* (Classe 1), e 29,65% das ausências e 70,35% das observações na classe *Chagas* (Classe 2). No que diz respeito à **Fração de Ejeção**, os valores mínimos, médias e máximos fornecem uma visão sobre a distribuição dessa variável. Observa-se uma faixa de valores que varia de 11,00 a 71,00, com uma média de

35,00 para ausências e 43,00 para observações.

Para a variável **AreaVDsis**, no caso do *Óbito*, 80,00% das ausências e 19,73% das observações são associadas a *Não óbito*, enquanto 93,10% das ausências e 6,90% das observações correspondem a *Óbito*. Quanto à **Classe Funcional** em relação à área do ventrículo direito, observam-se percentuais de 82,35% para ausências e 17,65% para observações na categoria *Assintomático* (Classe 1). As categorias *Levemente assintomático* (Classe 2), *Sintomático* (Classe 3) e *Sintomático em repouso*(Classe 4) apresentam percentuais de 85,39%, 85,42% e 92,98% para ausências, respectivamente, e 14,61%, 14,58% e 7,02% para observações. Finalmente, a Sorologia Chagas mostra 87,59% das ausências e 12,41% das observações na classe *Idiopático* (Classe 1), enquanto 84,42% das ausências e 15,58% das observações estão na classe *Chagas* (Classe 2).

Após a análise dos grupos observados e não observados nas variáveis, podemos levar em consideração a suposição de que os mecanismos envolvidos nas variáveis **Razão TEI** e **AreVDsis** podem ser caracterizados como MAR ou MNAR. A taxa de ausência varia entre as diferentes categorias e variáveis observadas, isso pode indicar que o mecanismo é MAR. Por exemplo, observar uma alta taxa de ausência em categorias como Óbito e Classe Funcional sugere que a probabilidade de ausência pode estar associada a essas variáveis observadas, como a presença de uma condição clínica ou um nível de gravidade dos sintomas.

Podemos levar em consideração a suposição de que os mecanismos de ausência nas variáveis **Razão TEI** e **AreaVDsis** podem ser caracterizamos como MAR ou MNAR. Por exemplo, observa-se uma alta taxa de ausência em categorias como **Óbito** e **Classe Funcional** indicando que a probabilidade de ausência pode estar associada a presença de uma condição clínica ou nível de gravidade dos sintomas e esta associação pode caracterizar MAR.

Por outro lado, a probabilidade de ausência de dados pode depender das variáveis que estão faltando, ou seja, o fato de um dado estar ausente está relacionado com o valor que ele teria se estivesse presente. Se a ausência de dados é maior em categorias específicas ou varia de forma com as características não observadas, pode sugerir um mecanismo MNAR.

Diante do problema de ausência no banco de dados e assumindo que se não tratarmos os dados estaremos perdendo informações relevantes, optamos por aplicar os métodos de correção via IM nas variáveis **Razão TEI** e **AreaVDsis**. Para isso consideramos os métodos: NORM, CONG, NA e CART. Consideramos $K = 5$ conjuntos de dados imputados. A análise CC foi considerada para fins de comparação. Além disso, após o tratamento dos dados via IM ajusta-se o modelo de taxas de falhas proporcionais.

Após o ajuste do modelo de taxas de falhas proporcionais e a obtenção das estimativas dos coeficientes, é possível calcular a Razão de Taxa de Falha (RTF) exponenciando os efeitos observados. Considerando um intervalo de confiança de 95% para a análise CC, as variáveis significativas ao nível de 5% foram: **Classe Funcional 3 - Sintomático**,

que apresentou uma RTF de 9,001. Isso significa que a taxa de óbito para indivíduos sintomáticos é 9,001 vezes maior que a taxa de óbito para indivíduos na **Classe Funcional 1 - Assintomático**. Outra variável significativa na análise CC foi a **Sorologia Chagas 2 - Chagas**, com uma RTF de 13,616, o que indica que indivíduos com sorologia positiva para Chagas apresentam uma taxa de óbito 13,616 vezes maior do que indivíduos do grupo **Sorologia Chagas 1 - Idiopático** (Tabela 14).

Tabela 14 – Resultados da aplicação prática

Métodos	Variável	Coef. (θ)	IC (Coef)	RTF	IC (RTF)	P-valor
1. CC	Classe Funcional 2	0,160	(-1,418; 1,738)	1,174	(0,242;5,687)	0,842
	Classe Funcional 3	**2,197**	**(0,166; 4,229)**	**9,001**	**(1,180;68,638)**	**0,034**
	Classe Funcional 4	2,220	(-0,104; 4,544)	9,210	(0,902;94,077)	0,061
	Sorologia Chagas 2	**2,611**	**(0,052; 5,170)**	**13,616**	**(1,054;175,932)**	**0,045**
	Fração de Ejeção	-0,006	(-0,091; 0,078)	0,994	(0,913;1,082)	0,884
	Razão TEI	2,020	(-0,413; 4,454)	7,540	(0,661;85,955)	0,104
	AreaVDsis	-0,022	(-0,137; 0,093)	0,978	(0,872;1,098)	0,709
2. NORM	Classe Funcional 2	0,364	(-0,042; 0,770)	1,439	(0,959;2,160)	0,078
	Classe Funcional 3	**0,619**	**(0,173; 1,065)**	**1,857**	**(1,189;2,902)**	**0,007**
	Classe Funcional 4	**1,104**	**(0,596; 1,611)**	**3,015**	**(1,815;5,010)**	**< 0,000**
	Sorologia Chagas 2	**0,780**	**(0,387; 1,172)**	**2,181**	**(1,473;3,230)**	**< 0,000**
	Fração de Ejeção	**-0,052**	**(-0,075; -0,030)**	**0,949**	**(0,928;0,971)**	**< 0,000**
	Razão TEI	**0,726**	**(0,018; 1,435)**	**2,068**	**(1,018;4,198)**	**0,045**
	AreaVDsis	0,036	(-0,039; 0,112)	1,037	(0,962;1,118)	0,263
3. CONG	Classe Funcional 2	0,253	(-0,103; 0,608)	1,288	(0,902;1,837)	0,163
	Classe Funcional 3	**0,569**	**(0,065; 1,073)**	**1,766**	**(1,067;2,923)**	**0,028**
	Classe Funcional 4	**1,188**	**(0,693; 1,682)**	**3,279**	**(2,000;5,378)**	**< 0,000**
	Sorologia Chagas 2	**0,908**	**(0,555; 1,262)**	**2,480**	**(1,742;3,532)**	**< 0,000**
	Fração de Ejeção	**-0,064**	**(-0,088; -0,041)**	**0,938**	**(0,916;0,960)**	**< 0,000**
	Razão TEI	**1,144**	**(0,444; 1,843)**	**3,138**	**(1,560;6,316)**	**0,003**
	AreaVDsis	-0,001	(-0,097; 0,095)	0,999	(0,908;1,100)	0,984
4. NA	Classe Funcional 2	0,310	(-0,084; 0,705)	1,364	(0,919;2,024)	0,121
	Classe Funcional 3	**0,624**	**(0,153; 1,095)**	**1,866**	**(1,165;2,989)**	**0,010**
	Classe Funcional 4	**1,114**	**(0,554; 1,674)**	**3,048**	**(1,741;5,336)**	**< 0,000**
	Sorologia Chagas 2	**0,771**	**(0,173; 1,370)**	**2,162**	**(1,189;3,934)**	**0,016**
	Fração de Ejeção	**-0,055**	**(-0,092; -0,018)**	**0,946**	**(0,912;0,982)**	**0,010**
	Razão TEI	0,912	(-0,425; 2,249)	2,489	(0,654;9,476)	0,144
	AreaVDsis	0,012	(-0,119; 0,143)	1,012	(0,888;1,154)	0,796
5. CART	Classe Funcional 2	0,256	(-0,100; 0,613)	1,292	(0,904;1,845)	0,158
	Classe Funcional 3	**0,593**	**(0,155; 1,031)**	**1,809**	**(1,168;2,803)**	**0,008**
	Classe Funcional 4	**1,193**	**(0,693; 1,693)**	**3,296**	**(1,999;5,435)**	**< 0,000**
	Sorologia Chagas 2	**0,861**	**(0,514; 1,208)**	**2,365**	**(1,672;3,346)**	**< 0,000**
	Fração de Ejeção	**-0,061**	**(-0,078; -0,045)**	**0,941**	**(0,925;0,956)**	**< 0,000**
	Razão TEI	**0,893**	**(0,271; 1,515)**	**2,442**	**(1,311;4,550)**	**0,007**
	AreaVDsis	0,007	(-0,021; 0,036)	1,007	(0,979;1,037)	0,599

Nota: As variáveis destacadas em negrito são significativas ao nível de 5%.
Coef.: Coeficientes; **RTF**: Razão Taxa de Falha; **IC**: Intervalo de Confiança

Com 95% de confiança e considerando o modelo NORM, as variáveis significativas ao nível de 5% foram: **Classe Funcional 3 - Sintomático**, **Classe Funcional 4 - Sintomático em repouso**, **Sorologia Chagas 2 - Chagas**, **Fração de Ejeção** e **Razão TEI**. Ao considerar a RTF com a variável explicativa **Classe Funcional**, indivíduos sintomáticos apresentam uma taxa de óbito 1,857 vezes maior do que a de indivíduos assintomáticos. Além disso, indivíduos sintomáticos em repouso possuem uma taxa de

óbito 3,015 vezes superior à de indivíduos assintomáticos. No que se refere à variável **Sorologia Chagas**, pacientes com Chagas apresentam uma taxa de óbito 2,181 vezes maior do que a de pacientes idiopáticos. Para a variável **Fração de Ejeção**, a taxa de óbito é de 0,949, o que significa que, para um aumento de uma unidade na variável, o risco de óbito diminui em 5,1%. Quanto à **Razão TEI**, para cada unidade de aumento nessa variável, a taxa de óbito é aproximadamente 2,068 vezes maior, assumindo as demais variáveis constantes. Este valor sugere uma associação positiva entre a variável e o risco de evento.

Fundamentando-se nos resultados via imputação CONG, com 95% de confiança ao nível de 5% de significância, as variáveis explicativas foram as mesmas que no modelo NORM. Todavia, os efeitos dos parâmetros são distintos, desta forma as interpretações para taxas de falhas são diferentes. Analisando a variável **Classe Funcional 3 - Sintomático**, pacientes sintomáticos possuem a taxa de óbito é de 1,766 vezes maior que dos pacientes sintomáticos. Observando a variável **Classe Funcional 4 - Sintomático em repouso** a taxa de óbito é de 3,279 vezes maior que a de indivíduos sintomáticos. Para **Sorologia Chagas 2 - Chagas** a taxa de óbito é de 2,480 vezes maior que indivíduos idiopáticos. Para a variável **Fração de Ejeção** a taxa de óbito é de 0,938, ou seja, para um aumento de uma unidade na variável a taxa de óbito diminui em 6,2%. Em relação à **Razão TEI**, que apresentou taxa de óbito igual à 3,138, indicando que para cada unidade de aumento na variável, a taxa de óbito é aproximadamente 3,138, assumindo as demais variáveis fixas, indicando uma associação positiva entre a variável e o risco de óbito.

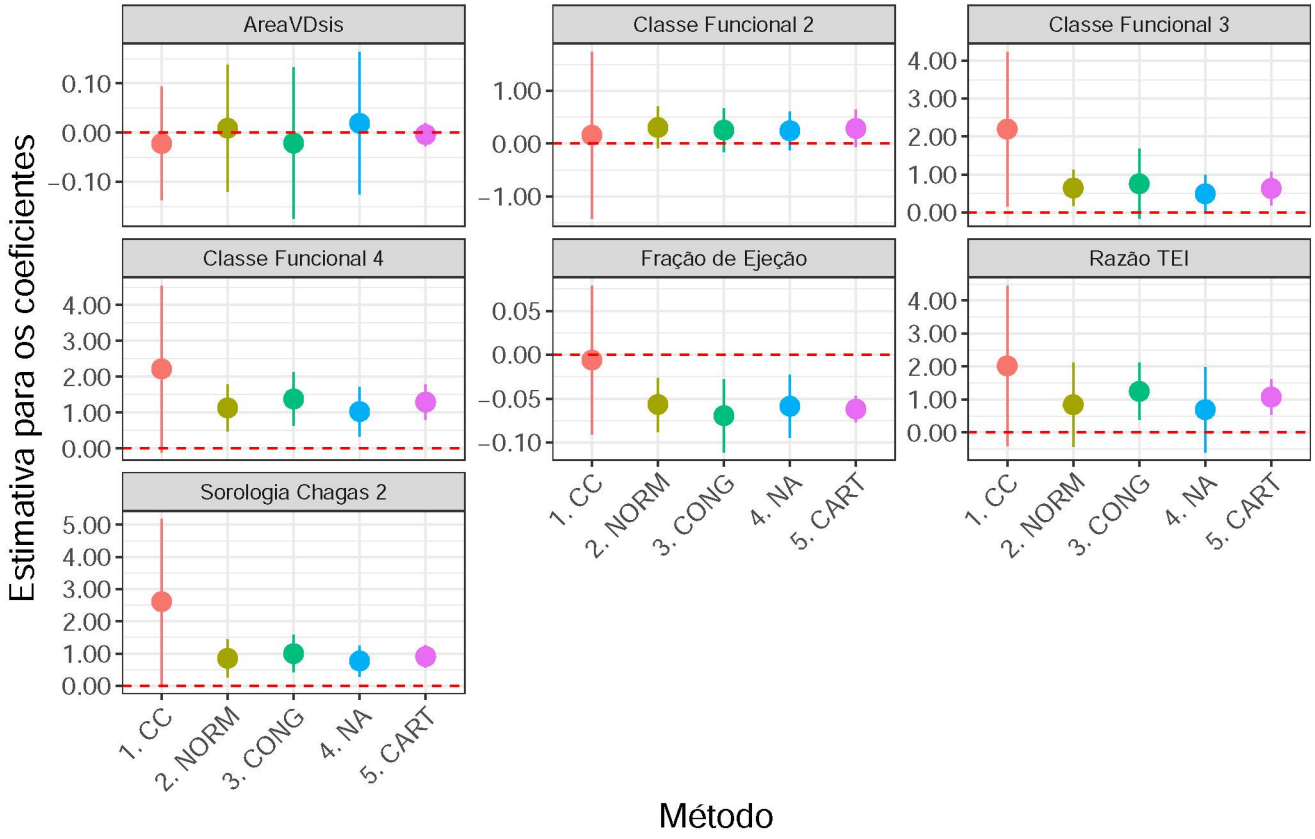
Para o modelo via NA, a variável **Classe Funcional 3 - Sintomático** apresentou taxa de óbito de 1,866 vezes maior que para indivíduos assintomáticos. A **Classe Funcional 4 - Sintomático em repouso** resultou em uma estimativa de taxa de óbito de 3,048 vezes maior que indivíduos sintomáticos. Analisando a taxa de óbito de **Sorologia Chagas 2 - Chagas** a taxa de óbito é de 2,162 vezes maior que pacientes idiopáticos. Para **Fração de Ejeção** o risco de óbito diminui em 5,4%.

Com base na imputação via CART, a **Classe Funcional 3 - Sintomático** a taxa de óbito é de 1,809 vezes maior que pacientes sintomáticos. Considerando a variável **Classe Funcional 4 - Sintomático em repouso** a taxa de óbito é de 3,296 vezes maior que indivíduos sintomáticos. Avaliando **Sorologia Chagas 2 - Chagas** a taxa de óbito é de 2,365 vezes maior que a de indivíduos idiopáticos. Para a **Fração de Ejeção** a razão taxa de falha é de 0,941, isto é, o risco de óbito diminui em 5,9%. Finalmente, para **Razão TEI** a taxa de óbito é de 2,442, assumindo as demais variáveis fixas, apontando uma associação positiva entre a variável e o risco de evento.

Todas as comparações da variável **Classe Funcional** foram realizadas em relação à categoria de referência sintomático. Todavia, podemos realizar a comparação entre grupos que é feita exponenciando a diferença entre coeficientes. Por exemplo, se quisermos comparar **Classe Funcional 4 - Sintomático em repouso** e **Classe Funcional 3 -**

Sintomático para o modelo via NORM o valor da RTF foi de 0,485. O IC para a RTF resultante pode ser obtido realizando a diferença entre os limites dos coeficientes e depois exponenciando os valores, desta forma teremos (0,420; 0.546).

Figura 1 – IC para estimativas da RTF por método



A Figura 1 representa os intervalos de confiança para a RTF por método. Os intervalos foram colocados em escala logarítmica para melhorar a visualização, pois a análise CC retornou intervalos de confiança amplos o que significa maior incerteza na estimativa dos parâmetros e reforça os problemas que podem ser gerados neste tipo de análise. Como estamos em escala logarítmica a RTF's a linha tracejada em 0 representa o valor 1 na escala original. Considerando os métodos de imputação, podemos observar que as variáveis que sofreram imputação, **AreaVDsis** e **Razão TEI** possuem intervalos de confiança mais largos dependendo do método. Por exemplo, o método NORM apresentou intervalos largos, enquanto o CART apresentou intervalos mais estreitos. Para as variáveis que não sofreram imputação os intervalos ficam próximos para **Classe Funcional 2**, **Classe Funcional 3** e **Sorologia Chagas 2**. Nota-se pequenas variações nos intervalos entre os métodos para as variáveis **Fração de Ejeção** e **Classe Funcional 4**.

6 CONCLUSÃO

Neste estudo, investigamos o método de imputação múltipla no modelo de taxas de falhas proporcionais, com o objetivo de analisar o comportamento das estimativas dos parâmetros em função de diferentes mecanismos de ausência e métodos de imputação. Foram aplicados desde modelos mais simples, como a regressão linear, até modelos mais complexos, como CART e CONG.

No estudo de simulação a análise CC e os modelos usuais de imputação (NORM e LOG) podem levar a resultados tendenciosos. Isto se deve ao fato que no CC estamos fazendo inferências para uma população diferente daquela que originou os dados. Os modelos de imputação usuais (NORM e LOG), por outro lado, são viesados por conta da natureza linear do modelo de imputação, o qual leva em conta uma distribuição condicional da variável ausente dado as demais. Por outro lado, os métodos CART e NA têm estimativas menos viesadas, mas subestimam a variância, resultando em baixas taxas de cobertura.

De forma geral, concluímos que o CONG se destacou em quase todos os cenários, mostrando-se o mais robusto e confiável. Este método apresentou vieses mínimos e taxas de cobertura próximas aos níveis nominais, tanto para variáveis contínuas quanto binárias, evidenciando sua eficácia em lidar com dados ausentes de forma geral. Os resultados obtidos indicam a importância crítica da escolha do método de imputação, pois pode impactar substancialmente a qualidade das estimativas e a interpretação dos resultados.

Na análise de dados reais, os modelos de imputação e a análise CC revelam diferenças importantes na significância e nas conclusões para as $RTF's$, decorrentes de aplicar um ou outro método. Desta forma, as inferências também serão distintas. Fundamentando-se no estudo de simulação podemos optar por interpretar o modelo de riscos proporcionais de Cox com correção via CONG. Pois este método apresentou menores valores de viés considerando mecanismos de perda como MAR e MNAR e perda em múltiplas variáveis. Além disso, o método apresentou taxas de cobertura próximas dos níveis nominais. Sugerindo que seria o modelo mais preciso que a análise CC e os demais métodos aplicados aos dados.

A aplicação desses métodos a dados reais ilustrou que a escolha do método de imputação impacta substancialmente as inferências e conclusões, sendo o CONG o mais adequado para correção de análises de dados com mecanismos de perda MAR e MNAR. Por fim, recomenda-se que estudos futuros explorem métodos alternativos, como método da máxima verossimilhança, bem como modelos mais complexos que considerem interações entre variáveis e diferentes taxas de censura, que podem influenciar a imputação e o desempenho dos modelos.

REFERÊNCIAS

AALEN, O. Nonparametric inference for a family of counting processes. **The Annals of Statistics**, v. 6, p. 701–726, 1978. 15

AUSTIN, P. C.; WHITE, I. R.; LEE, D. S.; BUUREN, S. v. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. **Canadian Journal of Cardiology**, 2020. 11

BARTLETT, J. W.; SEAMAN, S.; WHITE, I. R.; CARPENTER, J. R. Accommodating the model of interest within the fully conditional specification multiple imputation framework. 2012. 25, 26

BARTLETT, J. W.; SEAMAN, S. R.; WHITE, I. R.; CARPENTER, J. R. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive mode. **Statistical methods in medical research**, 2015. 5, 6, 12, 25

BOX-STEFFENSMEIER, J. M.; JONES, B. S. **Event History Modeling A Guide for Social Scientists**. [S.l.]: Cambridge University Press, 2004. 16

BREIMAN, L. **Classification and Regression Trees**. [S.l.]: Rotledge, 1984. v. 1. 28

_____. Random forests. **Kluwer Academic Publishers. Manufactured in The Netherlands**, 2001. 12

BURGETTE, L. F.; REITER, J. P. Multiple imputation for missing data via sequential regression trees. **American Journal of Epidemiology**, 2010. 29

CARPENTER, J. R.; KENWARD, M. G. **Multiple Imputation and its Application**. [S.l.]: John Wiley & Sons, Ltd, 2012. 11, 18, 19, 22, 23, 25, 26

CARROLL, O. U.; MORRIS, T. P.; KEOGH, R. H. How are missing data in covariates handled in observational time-to-event studies in oncology? a systematic review. **BMC Medical Research Methodology**, BioMed Central, v. 20, n. 1, p. 134, 2020. Disponível em: <https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01018-7>. 12

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. [S.l.]: Blucher, 2006. 14, 15, 17

COX, D. R. Regression Models and Life-Tables. **Imperial College, London**, 1972. 5, 6, 16

FALCARO, M.; NUR, U.; BERNARD, R.; CARPENTER, J. R. Estimating excess hazard ratios and net survival when covariate data are missing strategies for multiple imputation. **Epidemiology**, 2015. 17

GORETZKO, D.; HEUMANN, C.; BüHNER, M. Investigating Parallel Analysis in the Context of Missing Data: A Simulation Study Comparing Six Missing Data Methods. **Educational and Psychological Measurement**, 2019. 21

GUO, C.-Y.; YANG, Y.-C.; YI-HAU, C. The optimal machine learning-based missing data imputation for the cox proportional hazard model. **Frontiers in Public Health**, 2021. 12, 17

HSU, C.-H.; YU, M. Cox regression analysis with missing covariates via nonparametric multiple imputation. **Stat Methods Med Res.**, 2019. 12

JR., R. G. M.; GONG, G.; MUÑOZ, A. **Survival Analysis**. [S.l.]: A Wiley-Interscience Publication, 1998. 14

KLEIN, J. P.; MOESCHBERGER, M. L. **Survival Analysis Techniques for Censored and Truncated Data**. [S.l.]: Springer, 2003. 16

KLEINBAUM, D. G.; KLEIN, M. **Survival Analysis A Self-Learning Text**. [S.l.]: Springer, 2005. 14, 16

LITTLE, R. J. A.; RUBIN, D. B. **Statistical Analysis with Missing Data**. [S.l.]: Wiley Series in Probability and Statistics, 2002. 18

LITTLE, R. J. A.; SHENCKER, N. **Handbook of Statistical Modeling for the Social and Behavioral Sciences**. [S.l.]: Gerhard Aminger, Cliford C. Clogg, and Michael E. Sobel. Plenum Press, New York, 1995. 11

MOLENBERGHS, G.; KENWARD, M. G. **Missing Data in Clinical Studies**. [S.l.]: Wiley, 2007. v. 1. 19

NELSON, W. Theory and Applications of Hazard Plotting for Censored Failure Data. **Taylor & Francis, Ltd. on behalf of American Statistical Association and American Society for Quality**, v. 4, 1972. 15

NUR, U.; SHACK, L. G.; RACHET, B.; CARPENTER, J. R.; COLEMAN, M. P. Modelling relative survival in the presence of incomplete data: a tutorial. **International Jornal of Epidemiology**, 2010. 11

PIGOTT, T. D. A review of methods for missing data. **Educational Research and Evaluation: An International Journal on Theory and Practice**, 2010. 17, 20

RAGHUNNATHAN, T. E. What do we do with missing data? Some options for analysis of inclomplete data. **Annu. Rev. Public. Health**, 2004. 11

RUBIN, D. B. **Multiple Imputation for Nonresponse in Surveys**. [S.l.]: John Wiley & Sons, Inc., 1987. 5, 6, 12, 22

SAINANI, K. L. Dealing with missing data. **PMR**, 2015. 11, 18, 20

SCHAFER, J. L.; OLSEN, M. K. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. **Multivariate Behavioral Research**, 1998. 21

SEAMAN, S. R.; WHITE, I. R. Review of inverse probability weighting for dealing with missing data. **Statistical Methods in Medical Research**, 2011. 20

SILVA, J. L. P. A comparison of multiple imputation methods for the analysis of survival data with outcome related missing covariate values. **Sigmae, Alfenas**, v. 12, p. 76–89, 2023. 12, 29

WHITE, I. R.; ROYSTON, P. Imputing missing covariate values for the Cox model. **Statistics in Medicine**, v. 28, p. 3618–3637, 2009. 5, 6, 12, 21, 27, 28

YI, Y.; YE, T.; YU, M.; SHAO, J. Cox regression with survival-time-dependent missing covariate values. **Biometrics: A JORNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY**, p. 460–471, 2009. 11, 12

ZHOU, X.-H.; ECKERT, G. J.; TIERNEY, W. M. Multiple imputation in public health research. **Statistics in Medicine**, 2001. 21

7 APÊNDICE

7.1 PSEUDOCÓDIGO PARA SIMULAÇÕES

Algorithmus 1: Simulação para Estimativas de Cox com Diferentes Métodos de Imputação

Input: Tamanho das amostras *sizes*, Número de repetições *total*, Valor verdadeiro do parâmetro θ_{real}

Output: Resultados das estimativas e métricas

foreach *tamanho de amostra n em sizes* **do**
  for $i \leftarrow 1$ ***to*** *total* **do**
    Gerar X_1 e X_2 da Normal padrão ou da Binomial(1,0.5);
    Gerar $Z \sim \text{Binomial}(1, 0.5)$;
    Calcular taxa de risco: $\lambda = \exp(\theta_{\text{real}} \cdot X1 + \theta_{\text{real}} \cdot X2 + \theta_{\text{real}} \cdot Z)$;
    Gerar tempos de evento $t \sim \text{Exp}(\lambda)$;
    Gerar tempos de censura $C \sim \text{Weibull}(\text{shape} = 0.5, \text{scale} = 1)$;
    Obter tempos observados $t_{\text{obs}} = \min(t, C)$;
    Gerar indicador de censura $\delta = \mathbb{I}(t \leq C)$;
    Calcular probabilidades de ausência;
    $prop1 = (1 + \exp(-t_{\text{obs}}))^{-1}$;
    $prop2 = (1 + \exp(-(t_{\text{obs}} + 2 \cdot (X1 - 0.5) + Z)))^{-1}$;
    Gerar indicadores de ausência $R1 \sim \text{Binomial}(1, prop1)$;
    $R2 \sim \text{Binomial}(1, prop2)$;
    Aplicar ausência;
    $X_{1\text{aus}} = \text{se } R1 == 0 \text{ então } NA \text{ senão } X1$;
    $X_{2\text{aus}} = \text{se } R2 == 0 \text{ então } NA \text{ senão } X2$;
    Criar data frames: $dados_{\text{comp}}$, $dados_{\text{cc}}$, $dados_{\text{imp\_usual}}$, $dados_{\text{imp\_cong}}$, $dados_{\text{cart}}$, $dados_{\text{NA}}$;
    Ajustar modelos: FULL, CC, NORM, CONG, CART, NA;
    begin
      Modelo FULL: Ajustar Cox proporcional com $dados_{\text{comp}}$;
      Modelo CC: Ajustar Cox proporcional com $dados_{\text{cc}}$;
      Modelo NORM: Imputar com método normal e ajustar Cox;
      Modelo CONG: Imputar com método congenial e ajustar Cox;
      Modelo CART: Imputar com método CART e ajustar Cox;
      Modelo NA: Imputar com método NA e ajustar Cox;
    Salvar resultados das estimativas e métricas;
    Erro Continuar com a próxima iteração;
Calcular métricas de avaliação a partir dos resultados;

7.2 FIGURAS COMPLETARES DA SIMULAÇÃO

Figura 2 – Curvas de sobrevivência via Kaplan-Meier

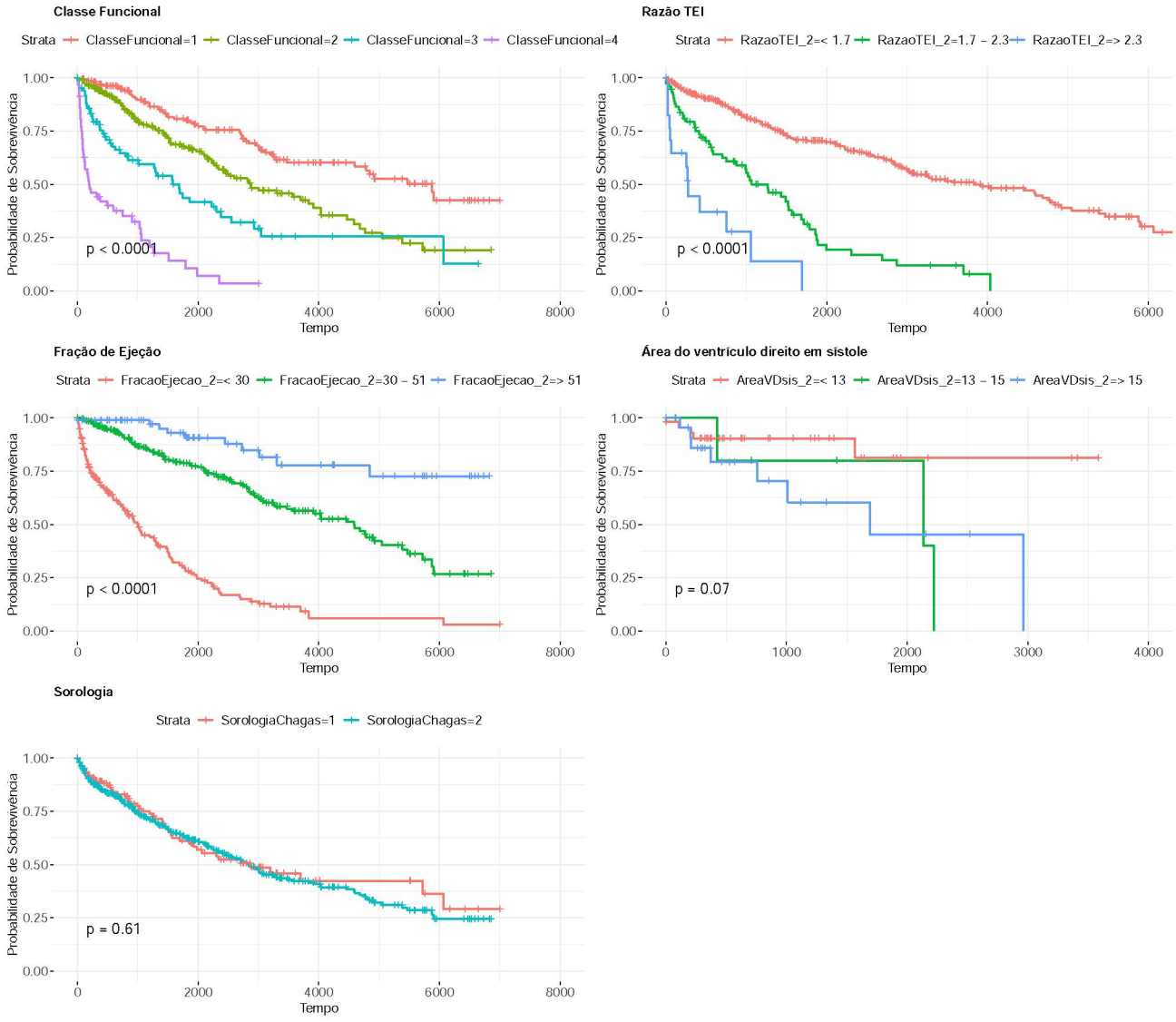


Figura 3 – Variáveis presentes na base de dados de Chagas

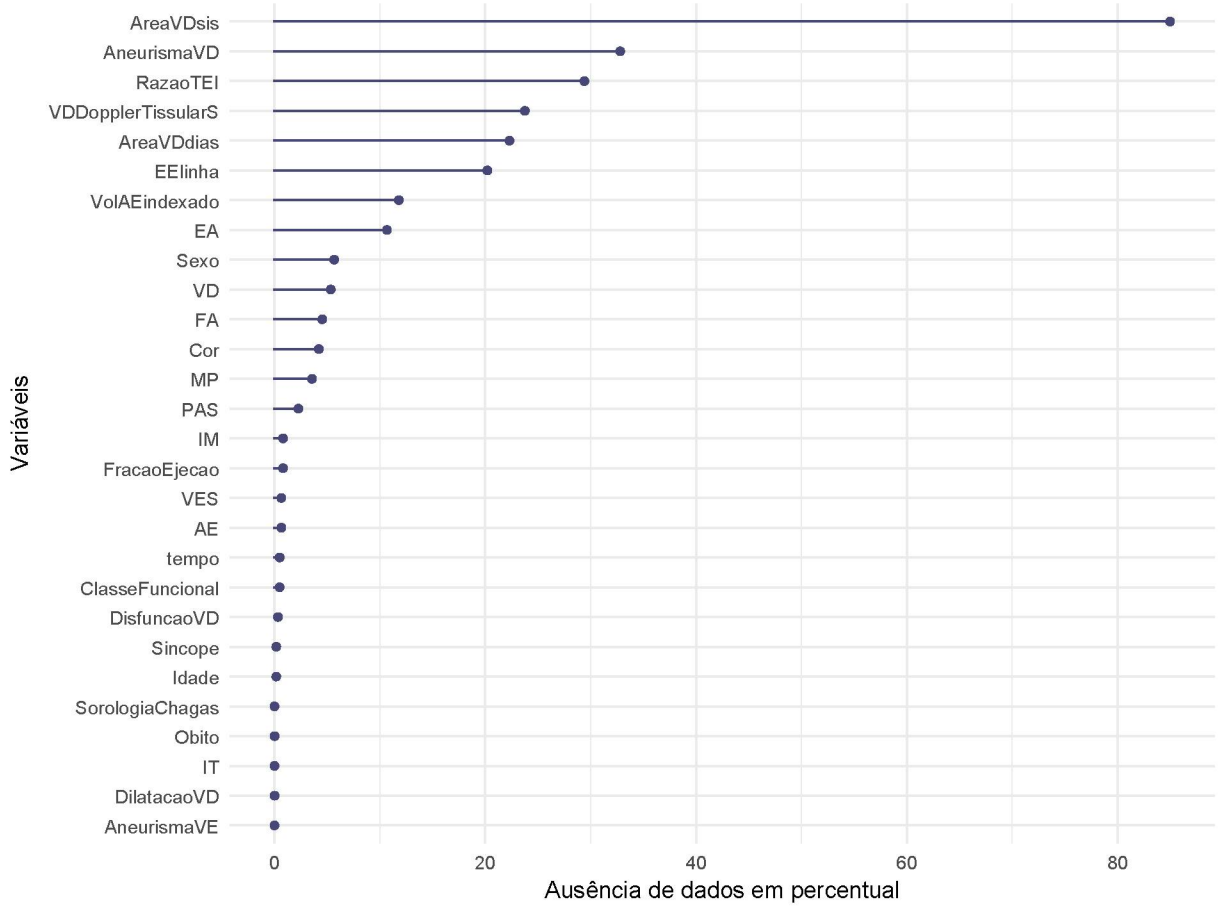
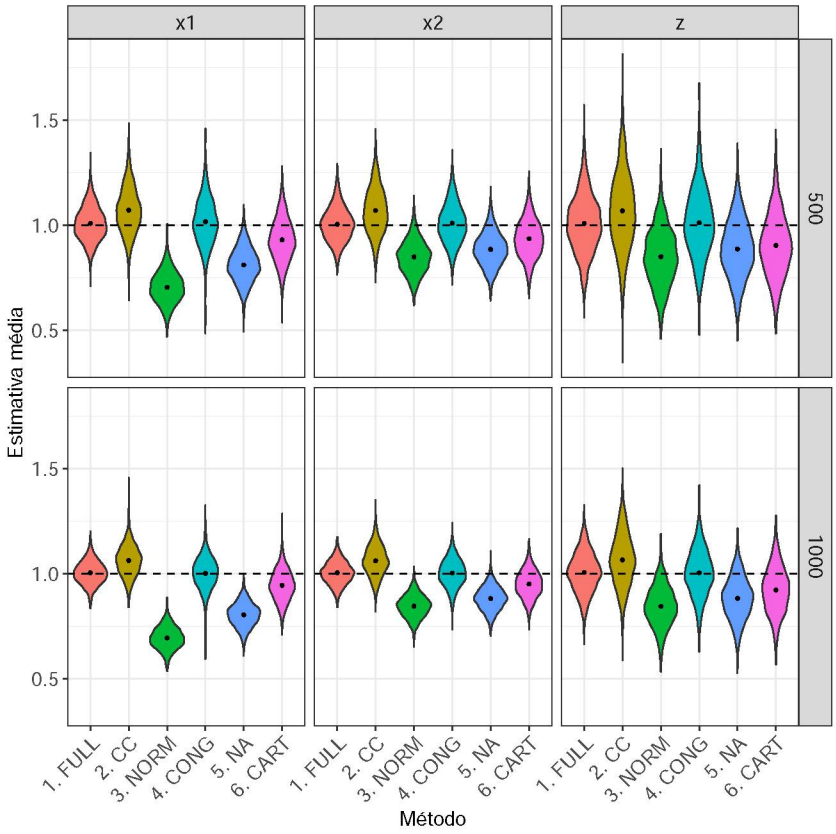
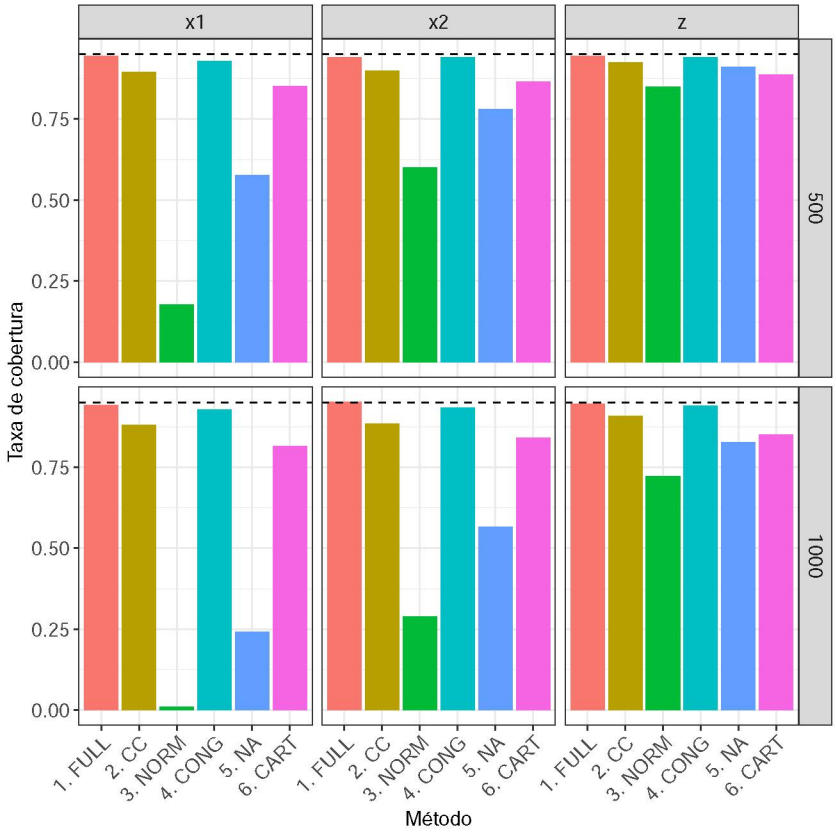


Figura 4 – Estimativa média dos parâmetros e taxas de cobertura para X_1 (MAR)

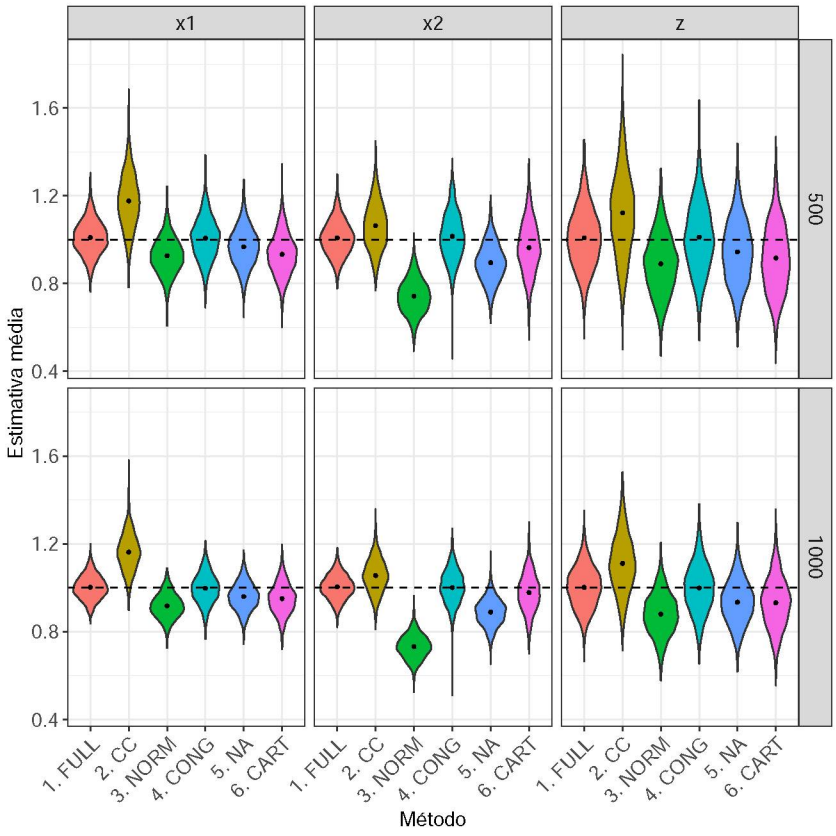


(a) Estimativa média dos parâmetros X_1 (MAR).

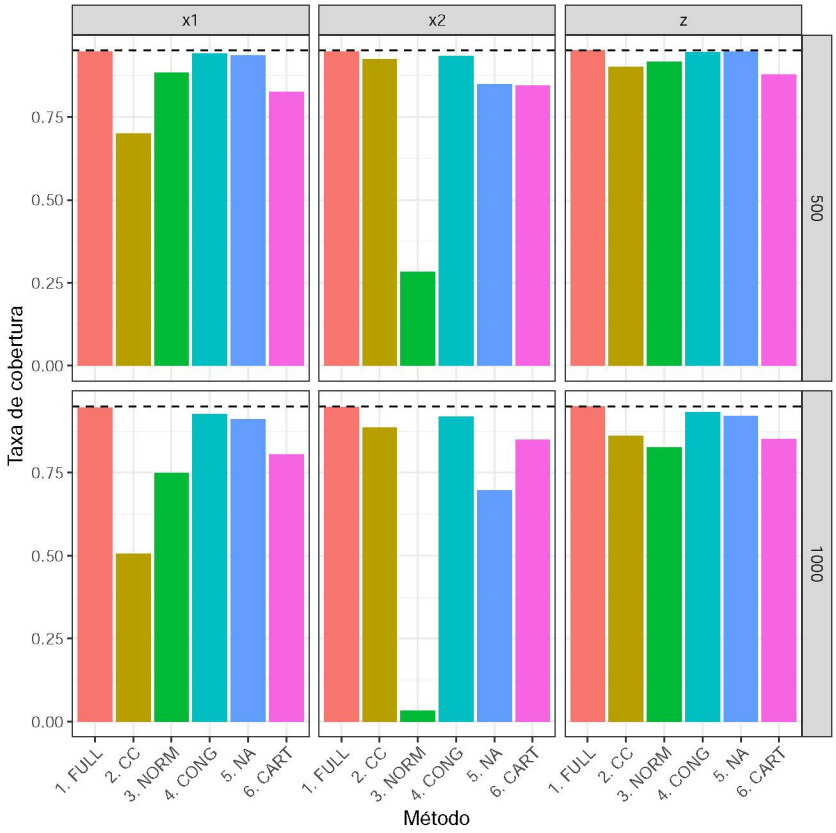


(b) Taxas de cobertura para X_1 (MAR).

Figura 5 – Estimativa média dos parâmetros e taxas de cobertura empírica para X_2 (MAR)

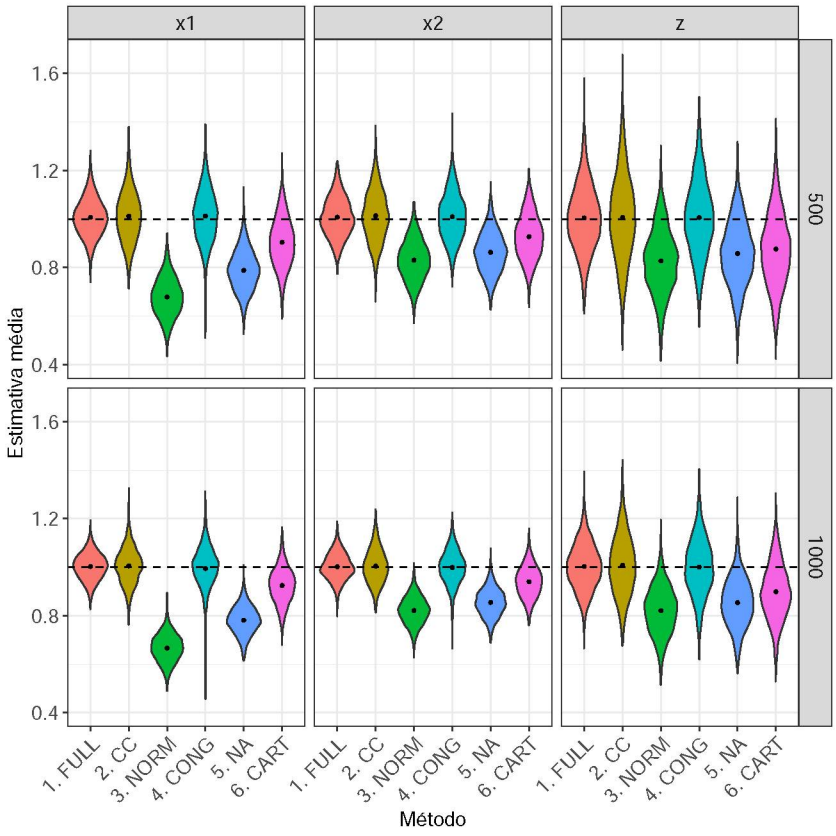


(a) Estimativa média dos parâmetros X_2 (MAR).

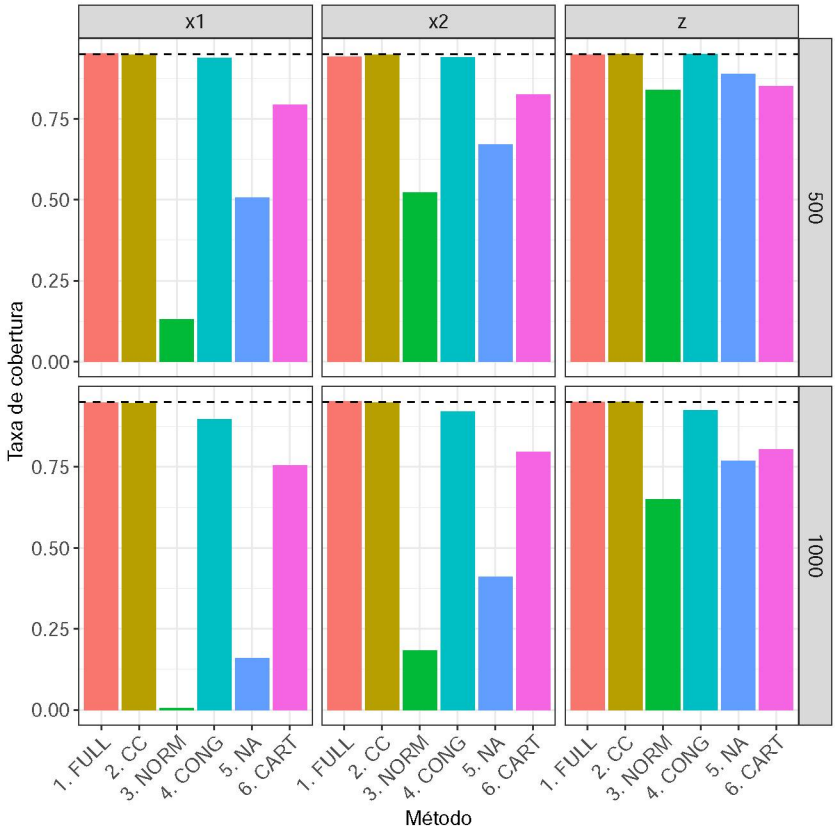


(b) Taxas de cobertura para X_2 (MAR).

Figura 6 – Estimativa média dos parâmetros e taxas de cobertura empírica para X_1 (MCAR)

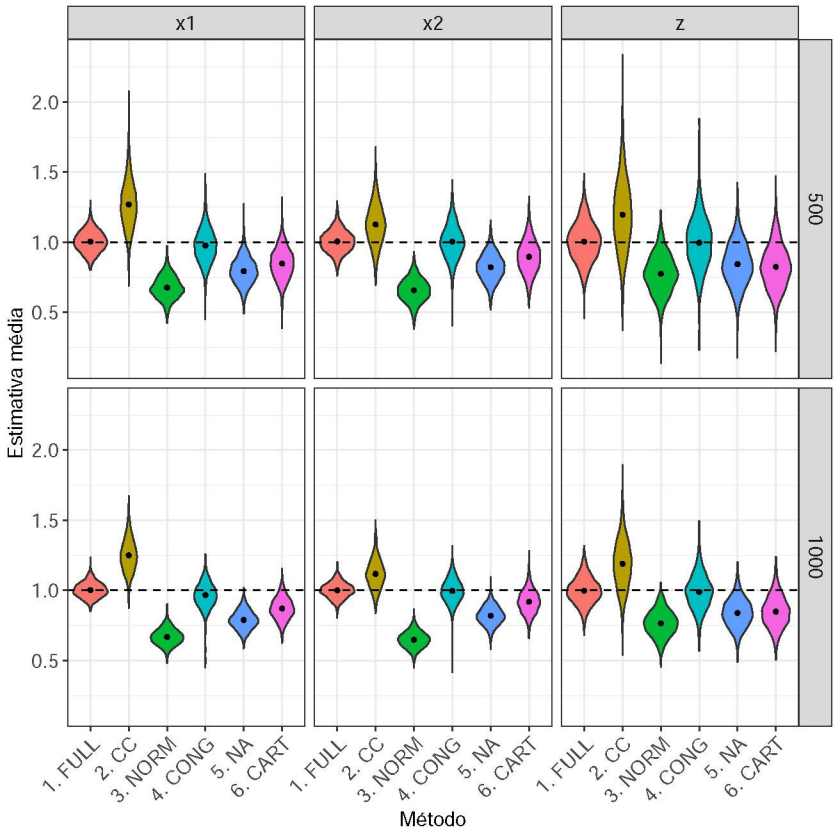


(a) Estimativa média dos parâmetros X_1 (MCAR).

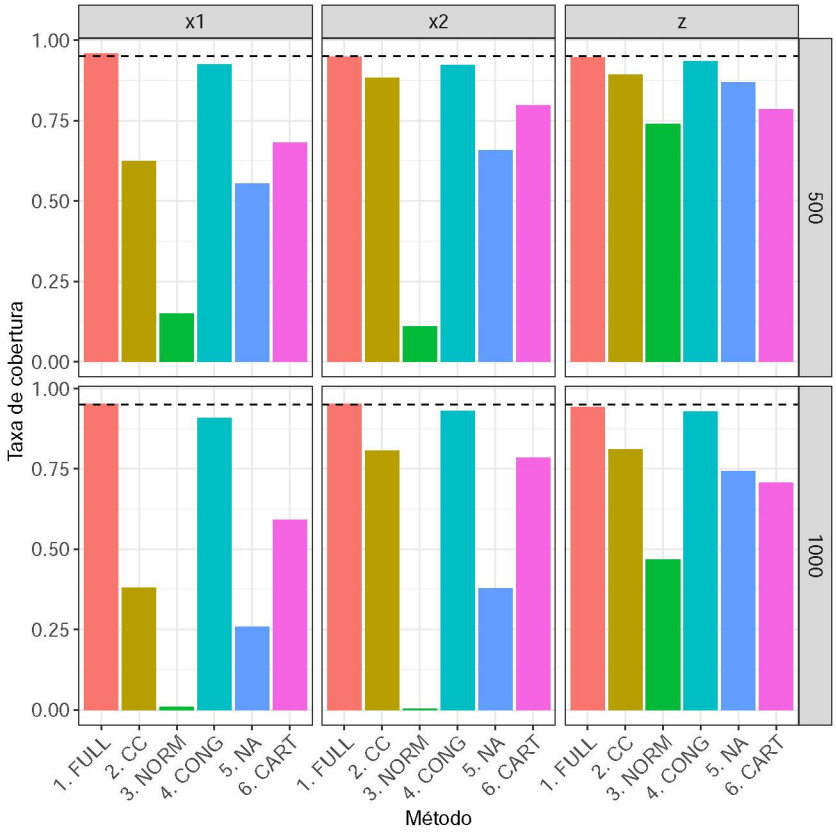


(b) Taxas de cobertura para X_1 (MCAR).

Figura 7 – Estimativa média dos parâmetros e taxas de cobertura empírica para X_1 (MAR) e X_2 (MNAR)

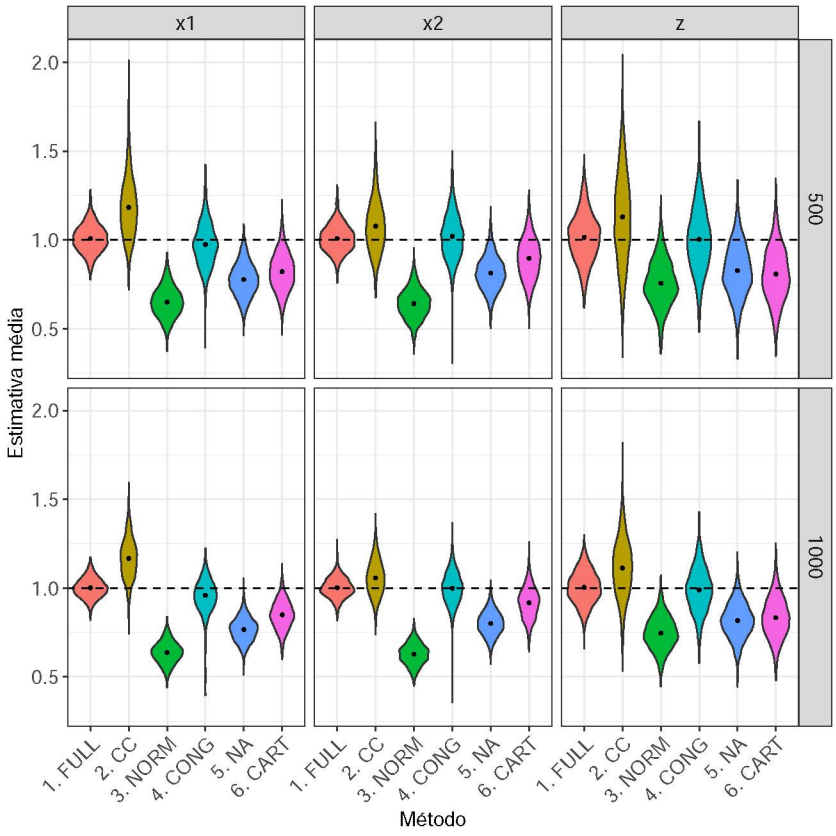


(a) Estimativa média dos parâmetros X_1 e X_2 (MAR MNAR).

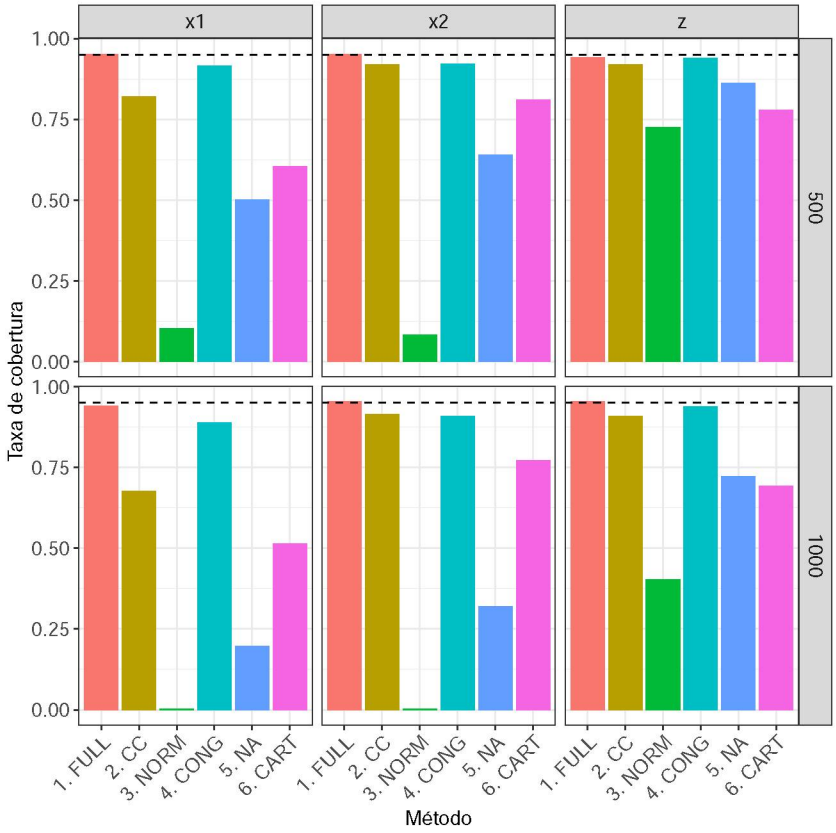


(b) Taxas de cobertura para X_1 e X_2 (MAR MNAR).

Figura 8 – Estimativa média dos parâmetros e taxas de cobertura empírica para X_1 (MCAR) e X_2 (MAR)

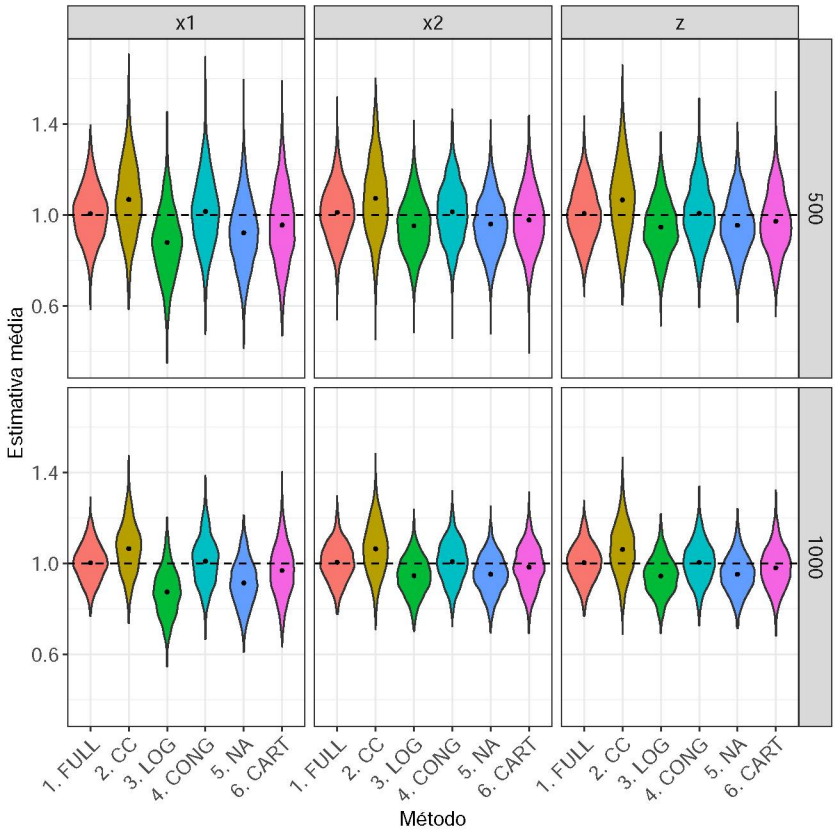


(a) Estimativa média dos parâmetros X_1 e X_2 (MCAR MNAR)

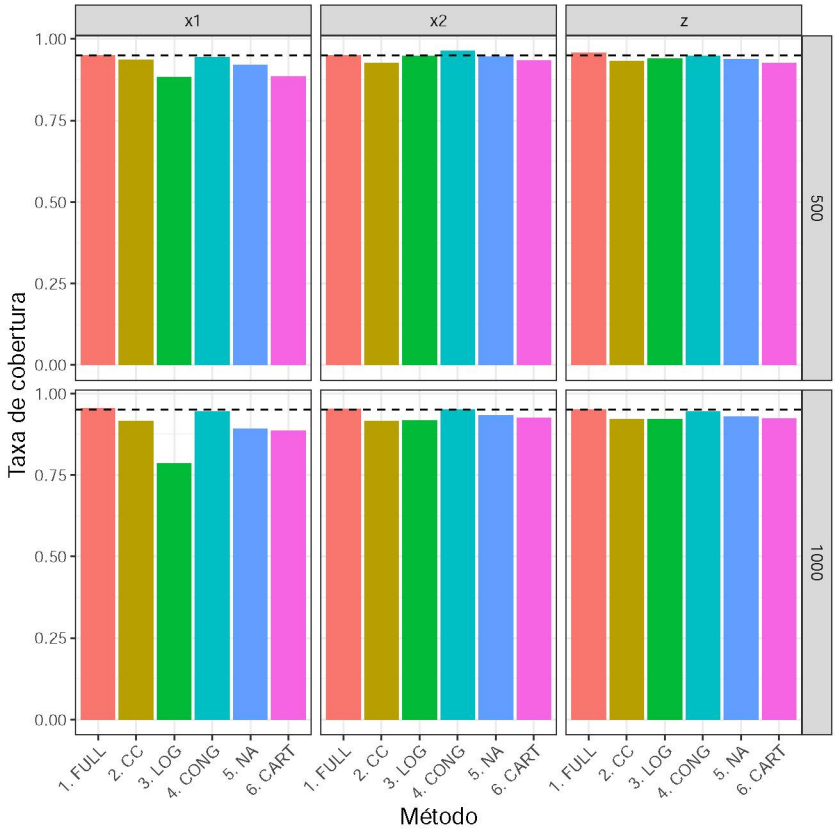


(b) Taxas de cobertura para X_1 e X_2 (MCAR MNAR).

Figura 9 – Estimativa média dos parâmetros e taxas de cobertura para X_1 (MAR)

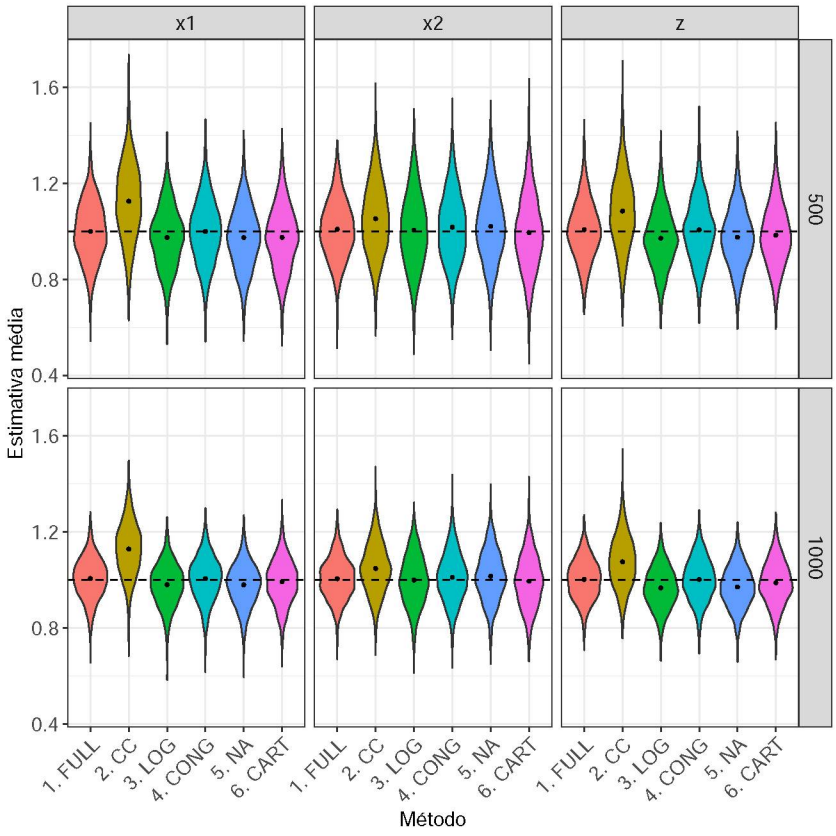


(a) Estimativa média dos parâmetros X_1 (MAR).

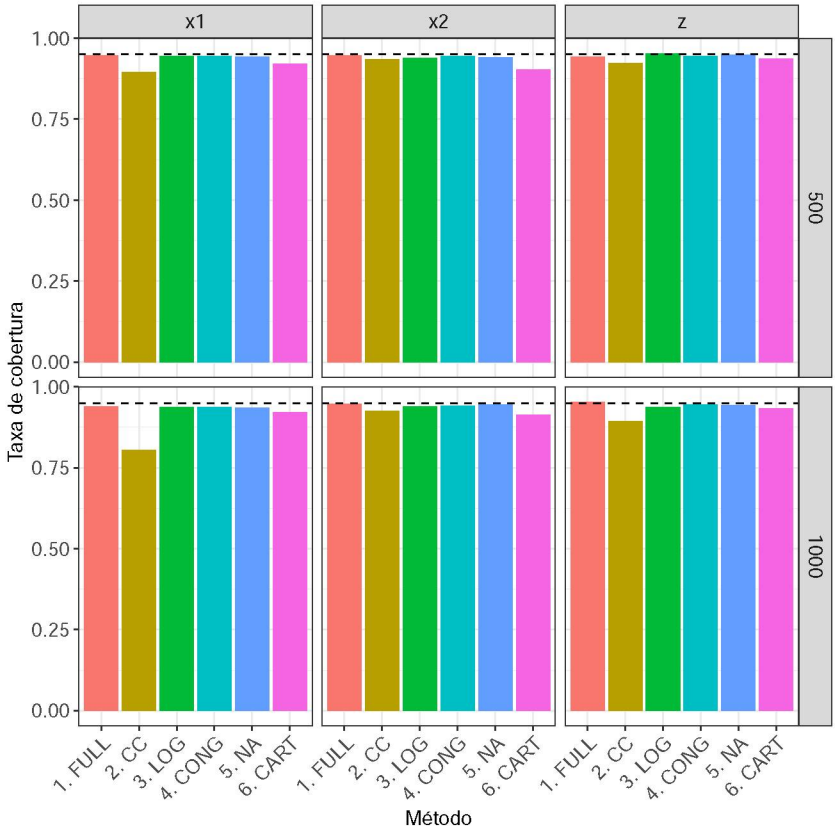


(b) Taxas de cobertura para X_1 (MAR).

Figura 10 – Estimativa média dos parâmetros e taxas de cobertura empírica para X_2 (MAR)

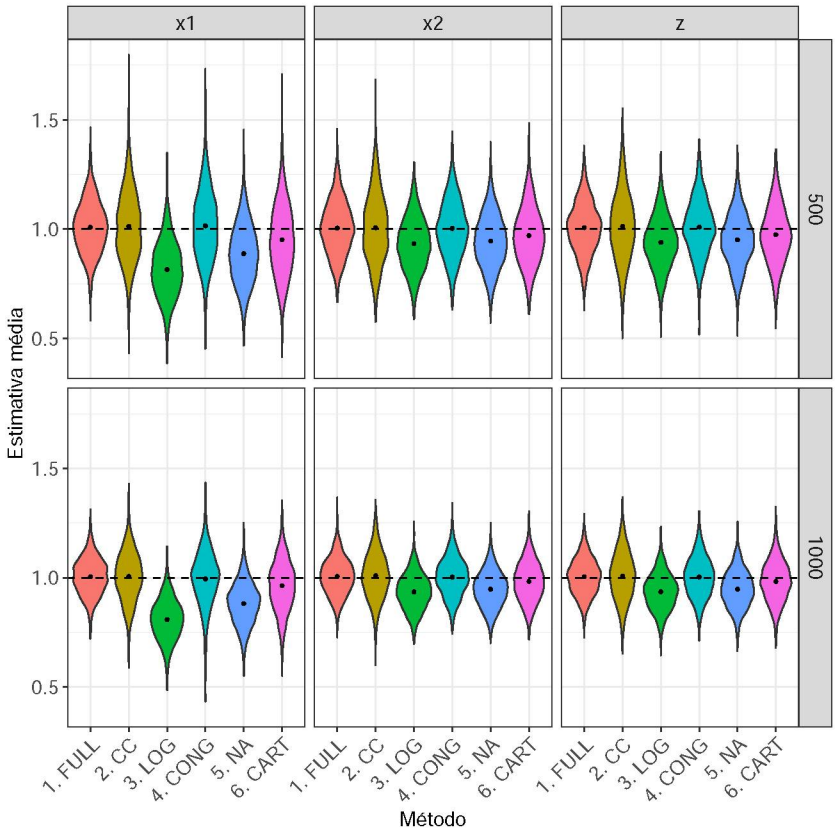


(a) Estimativa média em X_2 (MAR).

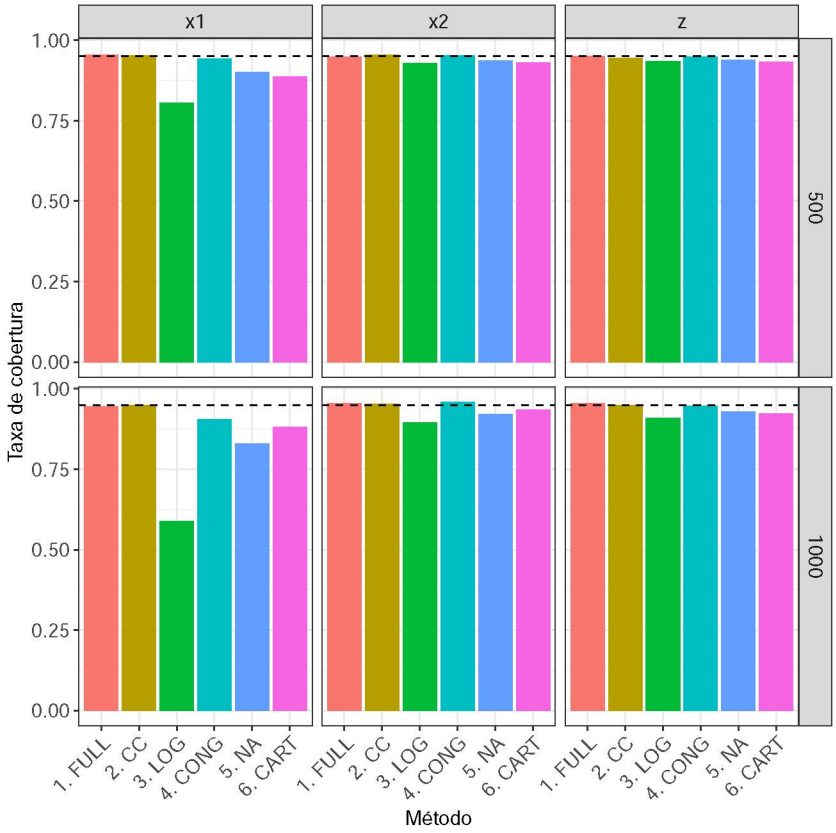


(b) Taxas de cobertura empírica para X_2 (MAR).

Figura 11 – Estimativa média dos parâmetros e taxas de cobertura empírica para X_1 (MCAR)

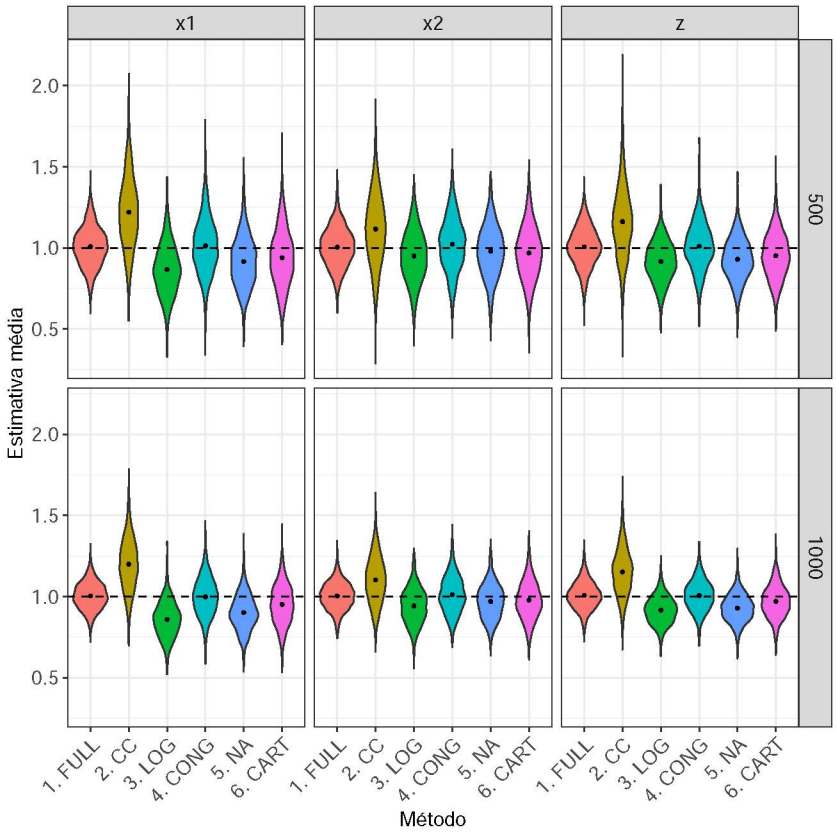


(a) Estimativa média para X_1 (MCAR).

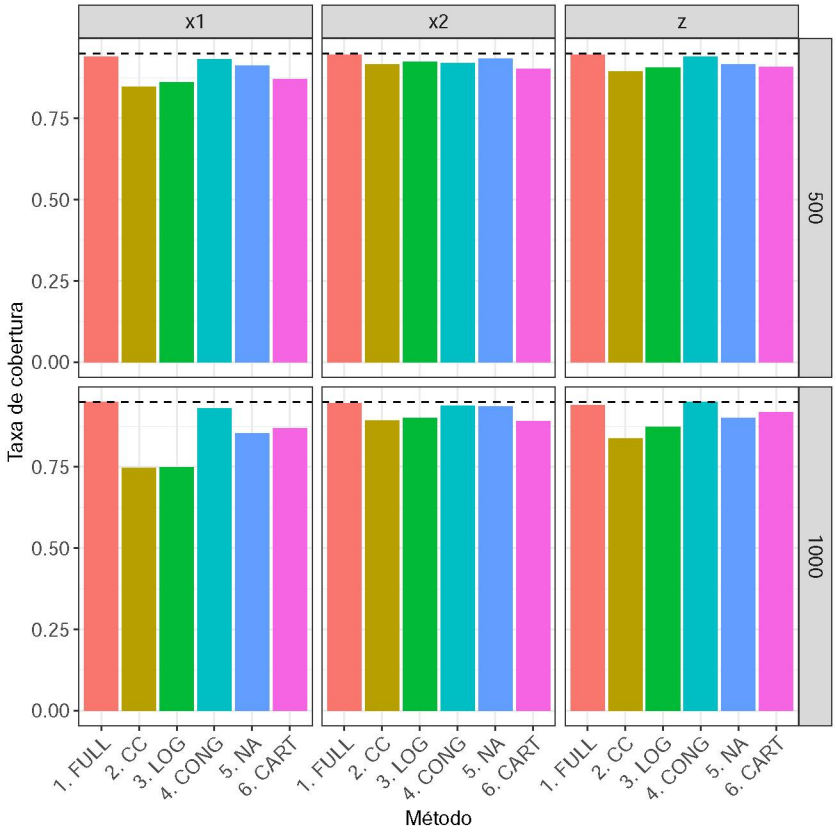


(b) Taxas de cobertura empírica para X_1 (MCAR).

Figura 12 – Estimativa média dos parâmetros e taxas de cobertura empírica para X_1 (MAR) e X_2 (MNAR)

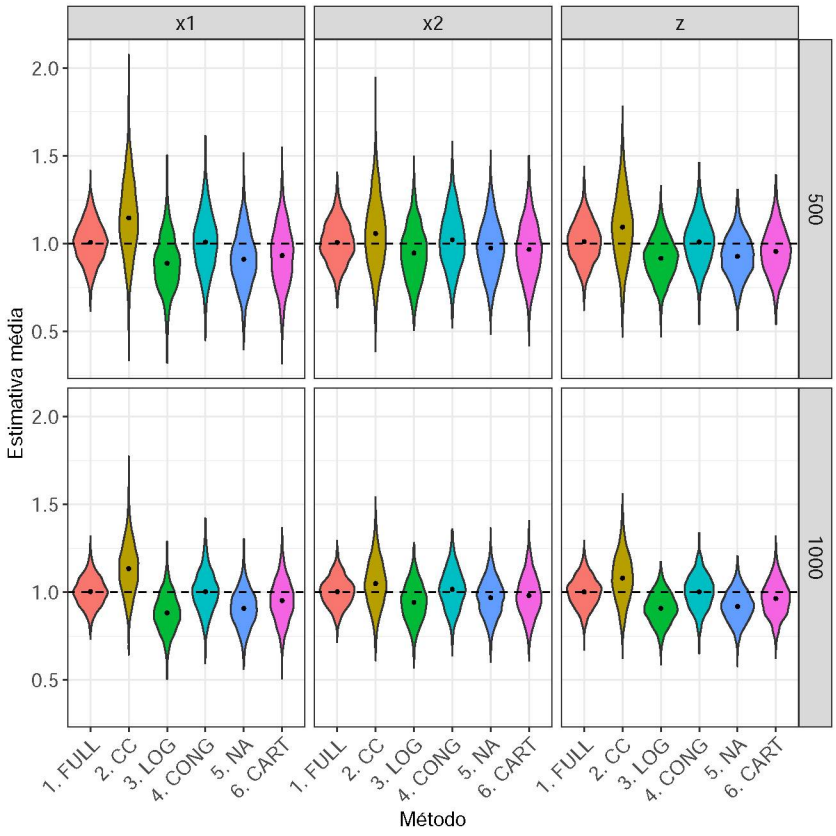


(a) Estimativa média para X_1 e X_2 (MAR, MNAR).

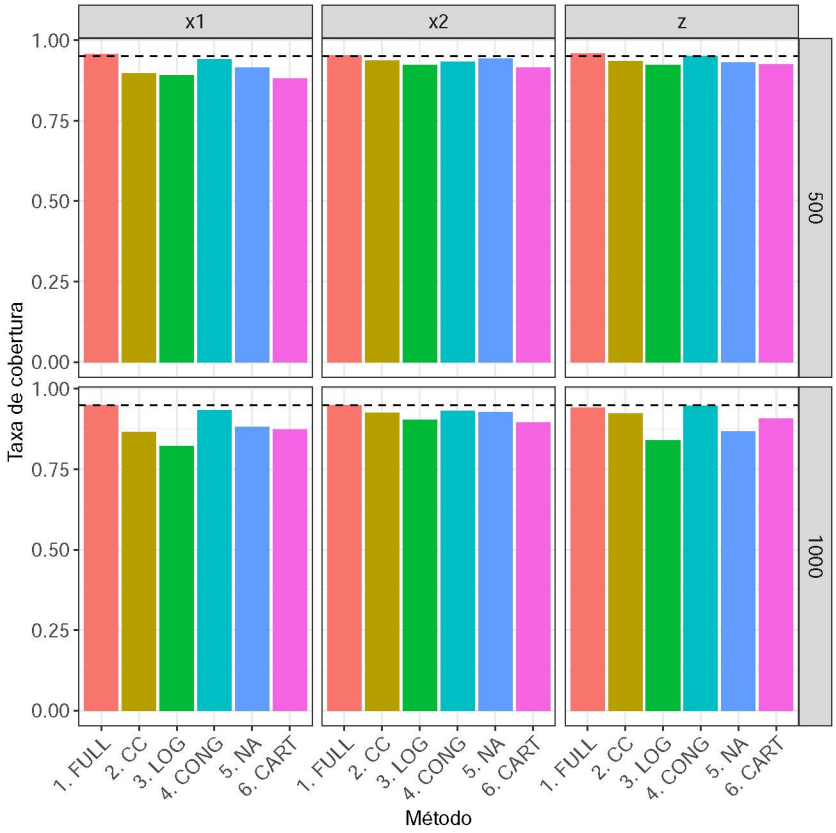


(b) Taxas de cobertura empírica para X_1 X_2 e (MAR, MNAR).

Figura 13 – Estimativa média dos parâmetros e taxas de cobertura empírica para X_1 (MCAR) e X_2 (MAR)



(a) Estimativa média para X_1 e X_2 (MCAR, MNAR).



(b) Taxas de cobertura empírica para X_1 X_2 e (MCAR, MNAR).